HiTZ-Ixa at SemEval-2025 Task 1: Multimodal Idiomatic Language Understanding

Anar Yeginbergen, Elisa Sanchez-Bayona, Andrea Jaunarena and Ander Salaberria
HiTZ Center, University of the Basque Country
{anar.yeginbergen, elisa.sanchez, andrea.jaunarena, ander.salaberria}@ehu.eus

Abstract

In this paper, we present our approach to the AdMIRe (Advancing Multimodal Idiomaticity Representation) shared task, outlining the methodologies and strategies employed to tackle the challenges of idiomatic expressions in multimodal contexts. We discuss both successful and unsuccessful approaches, including the use of models of varying sizes and experiments involving zero- and few-shot learning. Our final submission, based on a zero-shot instruction-following vision-and-language model (VLM), achieved 9th place for the English test set and 1st place for the Portuguese test set on the final leaderboard.

We investigate the performance of open VLMs in this task, demonstrating that both large language models (LLMs) and VLMs exhibit strong capabilities in identifying idiomatic expressions. However, we also identify significant limitations in both model types, including instability and a tendency to generate hallucinated content, which raises concerns about their reliability in interpreting figurative language. Our findings emphasize the need for further advancements in multimodal models to improve their robustness and mitigate these issues.

1 Introduction

While substantial progress has been made in the abilities of large language models (LLMs) to process literal meanings, their capacity to handle non-literal language remains an open research topic. The challenge is further amplified in multimodal settings, where figurative expressions may not have straightforward visual correspondences. Idioms, specifically, are typically defined as multi-word and non-compositional expressions. Non-compositionality implies that the meaning of the overall expression does not match the sum of meanings of its parts (words). Thus, they pose an inherent challenge for their understanding. For instance, an idiomatic expression like *"kick the bucket"* may not have a direct visual counterpart that aligns with its intended meaning in a sentence. Its use can be literal or idiomatic/figurative, requiring models to incorporate contextual and world knowledge for accurate interpretation.

In order to evaluate how current vision-and-language models (VLMs) align idiomatic and literal meanings to visual representations, Pickard et al. (2025) have built the AdMIRe (Advancing Multimodal Idiomaticity Representation) dataset and proposed a shared task with it. This shared task is composed of two subtasks where the ability to distinguish figurative and literal uses of idioms is needed to rank images by semantic similarity.

In our proposed approach, we investigate how well contemporary models handle figurative language in this multimodal context. By evaluating their performance on the first subtask of the AdMIRe dataset provided by the task organizers, we aim to assess whether these models can recognize and interpret idiomatic expressions effectively, moving beyond simple text-image correlations to capture deeper, non-literal meanings. Addressing this challenge is essential for advancing the interpretability and robustness of multimodal AI systems in real-world applications.

In this paper, we share our experiments and findings from our submission for the SemEval-2025 Task 1: Multimodal Idiomatic Language Understanding. We sum up our observations in our proposed solution to the shared task as follows:

- A state-of-the-art open-source vision-and-language model, *Qwen2-VL* (Wang et al., 2024), is capable of performing well using a zero-shot approach, achieving 73.3% and 100.0% accuracy on the small English and Portuguese test sets, respectively.
- *Qwen2-VL* struggles to do anything when we use a few-shot approach. We hypothesize that

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2256–2261
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

the model has not learned to process several sequences of images at the same time during pretraining, and that more sophisticated approaches are needed to solve the task in this manner.

- We show that, for this task, a smaller version of *Qwen2-VL* outperforms the biggest one, that is, the model with 7B parameters performs better than the 72B one.

2 Background

A key challenge in multimodal learning is ensuring that models can effectively capture both explicit and implicit relationships between visual and linguistic elements. While VLMs have shown remarkable progress in literal image-text alignment, their ability to understand figurative language, such as idioms and metaphors, remains an open research question. Idiomatic expressions often convey meanings that cannot be directly inferred from their constituent words, requiring models to move beyond surface-level associations and incorporate contextual understanding (Shwartz and Dagan, 2019).

Vision-Language Models (VLMs) have emerged as a powerful class of multimodal models that integrate visual and textual information, enabling machines to process and generate language in relation to images. These models build upon the success of large-scale pre-trained language models and vision encoders, leveraging architectures such as transformers (Vaswani et al., 2017) to bridge the gap between vision and language. By aligning textual and visual representations in a shared embedding space (Radford et al., 2021), VLMs have demonstrated impressive performance across various multimodal tasks, including image captioning (Anderson et al., 2018), visual question answering (Antol et al., 2015), and text-to-image generation (Ramesh et al., 2022).

Understanding figurative language requires models to go beyond literal word meanings and incorporate contextual, common sense, and world knowledge. Prior research has explored various approaches to tackling idioms, metaphors, and other forms of non-literal expressions, using both symbolic and neural methods.

Inside the figurative language we can find metaphors. The word **metaphor** was defined as a novel or poetic linguistic expression where one or more words for a concept are used outside of their normal conventional meaning to express a *similar* concept (Lakoff, 1993). So that, a linguistic metaphor takes a concept from a source domain and applies it to a target domain. Metaphors follow this schema: *TARGET* IS/ARE *SOURCE*. For example, behind the metaphor *She used some sharp words*, the source is *weapons* and the target is *words*, leading to *Words are weapons* relation. We can see another example with this metaphor: *I am the richest man in the world: I have the love of my family*, where the relation is *Well-being Is Wealth*, being *Wealth* the source and *Well-being* the target. As understanding metaphors with Large Language Models (LLMs) is something challenging, some previous works (Chakrabarty et al., 2023; Saakyan et al., 2024) have proposed working both hand in hand with linguistic metaphors and visual metaphors, in order to improve results in figurative language understanding. A visual metaphor is an image that wants to express a metaphorical message. Like in linguistic metaphors, visual metaphors also take a concept from a source domain and apply it to a target domain. Now, the main difference is that these domains need to be visually representable. Many experts have worked with linguistic metaphors; however, very few go deep into the multimodal field (Xu et al., 2024; Zhang et al., 2021).

Recent text-only approaches combine various types of figurative language, namely idioms, metaphor, sarcasm, or hyperbole, to improve models' understanding capabilities (Stowe et al., 2022; Lai et al., 2023; Kabra et al., 2023). Chakrabarty et al. (2021) introduced metaphor generation techniques using neural models, highlighting the potential of deep learning in handling figurative language. (Madabushi et al., 2021; Chakrabarty et al., 2022; Phelps et al., 2024; Liu et al., 2022) further analyzed the capabilities of current LLMs to understand idiomatic language.

3 Data

We are provided with the dataset for idiomatic expression understanding comprised from Madabushi et al. (2022), Pickard et al. (2025). Each idiomatic expression contains an idiomatic nominal compound (NC), a phrase where the NC is used in a literal or idiomatic way, and a set of 5 images.

The images for each sentence cover the following range of idiomaticity:

- A synonym for the idiomatic meaning of the NC.
- A synonym for the literal meaning of the NC.

Set	Language	#Examples	#Idiomatic	#Literal	#Avg. Words
Train	EN	70	39	31	124
Validation	EN	15	7	8	125
Test	EN	15	8	7	134
Test	EN_x	100	46	54	121
Train	PT	32	13	19	109
Validation	PT	10	5	5	112
Test	PT	13	6	7	114
Test	PT_x	55	31	24	108

Table 1: The distribution of the data. *EN* and *PT* are English and Portuguese data respectively. *EN_x* and *PT_x* pertain to the extended evaluation set.

- Something related to the idiomatic meaning, but not synonymous with it.
- Something related to the literal meaning, but not synonymous with it.
- A "distractor", which belongs to the same category as the compound (e.g. an object or activity) but is unrelated to both the literal and idiomatic meanings.

The goal of the tackled task was to rank the images according to the given sentence. Depending on the literal or idiomatic use of the NC in the sentence, the expected rank changes. Therefore, only a system that is capable of distinguishing such uses will be able to generalize well in this task. The overall distribution of the data is shown in Table 1.

4 System Description

In this section, we provide a description of our approach to *subtask A* of the shared task. Our primary objective is to investigate the capability of vision-and-language models (VLMs) to establish a meaningful connection between idiomatic expressions and their corresponding visual representations. Specifically, we aim to assess how effectively these models can associate idiomatic phrases with relevant visual cues and comprehend their intended meanings within a multimodal and multilingual context. By doing so, we seek to gain insights into the extent to which VLMs can interpret idiomatic language beyond literal meanings, leveraging both textual and visual information.

During the preliminary experiments, we discovered that the majority of the text-only LLMs are capable of recognizing idiomatic expressions from the context in which they occur no matter if it is idiomatic or literal. This finding motivated us to extend our investigation to VLMs under similar conditions. In our proposed solution, we employed *Qwen2-VL-7B-Instruct*[1] and *Qwen2-VL-72B-Instruct*[2] (Bai et al., 2023; Wang et al., 2024), by designing the instruction with the description of the task. This choice was taken due to two main reasons: i) its strong performance across different vision-and-language tasks and the capability of processing multiple images at the same time, which not many contemporary open-source VLMs can do.

Our approach is simple. We provided the details of the task in the prompt, explicitly indicating that one of the images is a distractor and that it should be placed in the last position of the output ranked images. The prompt we used in the experiments is illustrated in Figure 1.

5 Results

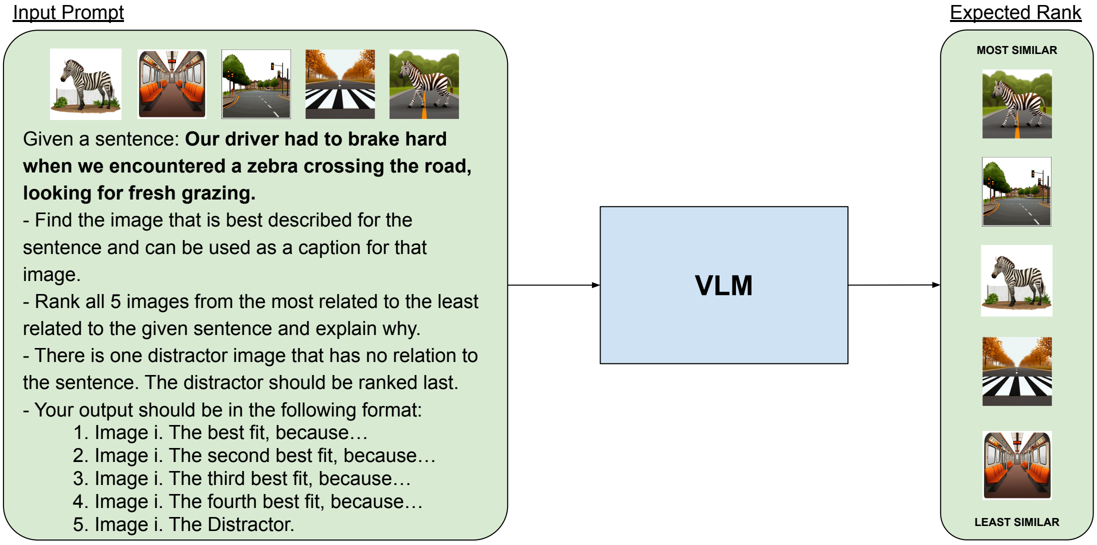
In this section, we will describe the results we obtained from different sets of experiments. We start with our main results obtained with the model described in Section 4, following with ablation studies carried out during our experimentation.

5.1 Main Results

In table 2, we show the performance and positions obtained in the test and extended test sets of *subtask A*. Top-1 accuracy measures the proportion in which the most similar images are ranked first in the output, whereas DCG measures the correct order of the entire ranking. For English, we ended up tied with 5 other participants in the 9th position of the leaderboard out of 33 contestants. For Portuguese, we achieved a perfect score and the 1st position in the test set due to a stroke of luck

[1] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[2] https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct



Given a sentence: **Our driver had to brake hard when we encountered a zebra crossing the road, looking for fresh grazing.**

- Find the image that is best described for the sentence and can be used as a caption for that image.
- Rank all 5 images from the most related to the least related to the given sentence and explain why.
- There is one distractor image that has no relation to the sentence. The distractor should be ranked last.
- Your output should be in the following format:
    1. Image i. The best fit, because…
    2. Image i. The second best fit, because…
    3. Image i. The third best fit, because…
    4. Image i. The fourth best fit, because…
    5. Image i. The Distractor.

Figure 1: Our system requires a VLM that can process multiple images at a time to output the rank of each input image with textual tokens, that is, following the format described in the input.

with the small number of test instances (see Table 1). The drop in the extended test set is a better representation of the performance that our naive approach has, which would leave us in the middle positions of the leaderboard.

If we combine the results obtained in both test and extended test sets, we get a top-1 accuracy of 60.0% and 58.9% for English and Portuguese, respectively. These combined sets contain 115 and 68 instances, which depict better the performance of our system. It is worth mentioning that the difference in performance across both languages is minimal. When comparing DCG metrics in these combined test sets, we end up with the same conclusion, as we get a DCG of 3.0 and 2.95 for English and Portuguese, respectively.

5.2 Ablation Studies

Model size. We have compared our system containing 7B parameters with the biggest *Qwen2-VL* model available, that is, *Qwen2-VL-72B-Instruct*. Even though models with a higher capacity usually show better performance across different tasks, this model fails to properly follow the provided instructions, being keen to hallucination.

On the contrary, we note that in our preliminary experiments, when we tried the text-only version to assess the quality of the LLMs for figurative language identification, *Llama-3.1-70B-Instruct* [3] performed better than *Llama-3.1-8B-Instruct*. [4]

Few-shot setting. We also investigated in-context few-shot learning by incorporating examples into the input prompt. However, due to limitations in context length and hardware, we were only able to include up to three examples from the training set. Despite this, the inclusion of these examples caused the Vision-Language Models (VLMs) to hallucinate, preventing the acquisition of reliable results. The issue was consistent across both 7B and 72B models.

In the technical report of *Qwen2-VL* family (Wang et al., 2024), the training and evaluation on vision-and-language tasks was done in a zero-shot manner. The model learned during training to intake several frames of a video at the same time, but it has never been trained with multiple sequences of images in a few-shot approach. Thus, we hypothesize that *Qwen2-VL* models are not capable of managing multiple sequences of multiple images in the same input prompt due to the lack of these examples during their multimodal training stage.

Object detection. In another approach, we wanted to analyze the effectiveness of object detectors in detecting figurative or literal usages of nominal compounds (NCs) in images. If the object detector were capable of detecting NCs only in

[3] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Position	Lang.	Team	Top-1 Acc.	DCG	Top-1 Acc. (Extended)	DCG (Extended)
9	EN	hitz_ixa	73.33	3.13	58.0	3.00
1	PT	hitz_ixa	100.00	3.51	45.45	2.82

Table 2: Position and performance in the preliminary leaderboard for both the English and Portuguese test sets and their respective extended test sets.

images where their use was literal, we would feed this information to the input prompt of *Qwen2-VL* models to ease their task.

By feeding the NC to an open-vocabulary object detector OwL-VIT with the template "A photo of a NC." (Minderer et al., 2022), we computed the proportions in which the NC is detected in different types of images.[5] On the one hand, the detection rate is 35% in images containing something related to the literal meaning of the NC. On the other hand, the ones relating a figurative meaning achieve a detection rate of 26%.[6] We conclude that this signal is too noisy to be useful and did not elaborate on further experimentation with object detectors. This noisiness can be easily understood when looking at Figure 1, where the figurative use of *zebra-crossing* (e.g. a crossing for pedestrians) has quite prominent visual features and can be easily detected by contemporary object detectors.

6 Conclusions

In this paper, we present our approach to the AdMIRe - Advancing Multimodal Idiomaticity Representation shared task, detailing the methodologies and strategies we employed to address the challenges posed by idiomatic expressions in multimodal contexts. We describe different successful and unsuccessful approaches, such as models of different size, zero- and few-shot experiments. With the final submission based on zero-shot instruction-following VLM, we were able to obtain 9th and 1st places in the preliminary leaderboard for English and Portuguese test sets respectively.

We explore the effectiveness of open vision-language models (VLMs) in achieving comparable performance in this task. While our findings indicate that LLMs demonstrate a strong ability to identify idiomatic expressions, and VLMs can achieve similar results, we also observe significant limitations in both model types. Specifically, these models often exhibit instability and are prone to generating hallucinated content, which raises concerns about their reliability in correctly interpreting figurative language. Our analysis highlights the need for future work for further improvements in multimodal models to enhance their robustness to mitigate these issues.

Acknowledgments

This work has been supported by the HiTZ Center, at the University of the Basque Country (UPV/EHU). Anar Yeginbergen's PhD contract is part of the PRE2022-105620 grant, financed by MCIN/AEI/10.13039/501100011033 and by the FSE+. Elisa Sanchez-Bayona is funded by a UPV/EHU grant "Formación de Personal Investigador". Ander Salaberria has received funding from the European Research Council under Horizon Europe, grant number 10113572, related to the LUMINOUS project.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative

[5] We use a threshold of 0.2 to discard low probability predictions.

[6] The detection rate for distractor images is 23%.

language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *Preprint*, arXiv:2305.14724.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff. 1993. *The contemporary theory of metaphor*, page 202–251. Cambridge University Press.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

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*.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *arXiv preprint arXiv:2109.04413*.

Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple open-vocabulary object detection with vision transformers. *ArXiv*, abs/2205.06230.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. *arXiv preprint arXiv:2405.09279*.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, Carolina Scarton, and Marco Idiart. 2025. AdMIRe: Advancing Multimodal Idiomaticity Representation (SemEval-2025 Task 1) - Labelled Datasets.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Understanding figurative meaning through explainable visual entailment. *Preprint*, arXiv:2405.01474.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multi-modal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.