YNU-HPCC at SemEval-2025 Task 1: Enhancing Multimodal Idiomaticity Representation via LoRA and Hybrid Loss Optimization

Lei Liu, You Zhang*, Jin Wang, Dan Xu, Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: liulei_academic@stu.ynu.edu.cn, yzhang0202@ynu.edu.cn

Abstract

This study reports the YNU-HPCC team's participation in Subtask A of SemEval-2025 Task 1 on multimodal idiomatic representation. The task requires ranking candidate images based on their semantic relevance to a target idiom within a given sentence, challenging models to disambiguate idiomatic semantics, and aligning them with abstract visual concepts across English and Portuguese. Using AltCLIP-m18 as the base model, our approach enhances its zero-shot capabilities with LoRA fine-tuning and combines ListMLE ranking optimization with Focal Loss to handle hard samples. Experimental results on the primary test set show significant improvements over the base model, with Top-1 Accuracy/DCG scores of 0.53/2.94 for English and 0.77/3.31 for Portuguese. The code is publicly available at https://github.com/1579364808/Semeval_2025_task1.

1 Introduction

Idioms, as a class of multiword expressions (MWEs), pose significant challenges for natural language understanding due to their non-compositional nature—their meanings cannot be derived from the literal interpretation of their constituent words (Dankers et al., 2022; Villavicencio et al., 2005). For instance, *bad apple* metaphorically refers to a disruptive individual rather than a decayed fruit. Despite the remarkable progress of pre-trained language models (PLMs) in text comprehension tasks, their ability to model idiomatic expressions remains limited. Key issues include susceptibility to literal meaning interference (Phelps et al., 2024; Chakrabarty et al., 2022; Madabushi et al., 2022) and insufficient grounding in multimodal experiences, such as visual perception (Lakoff and Johnson, 1980; Lu et al., 2023).

SemEval-2025 Task 1 introduces a multimodal evaluation framework, i.e., Advancing Multimodal Idiomaticity Representation (Pickard et al., 2025). Subtask A ranks candidate images based on their semantic relevance to a target idiom within a given sentence. This task requires models to disambiguate idiomatic semantics from textual contexts and align them with abstract visual concepts, presenting a significant challenge for current approaches. For example, in the *kangaroo court* case, the model must distinguish between the literal depiction of a kangaroo and the metaphorical representation of an unjust judicial process.

Given the task's bilingual nature (English and Portuguese), we propose a multilingual approach based on AltCLIP-m18 (Chen et al., 2022), a multilingual variant of the CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) model. We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient fine-tuning technique to efficiently adapt the model to the task. Additionally, we introduce a combined loss function integrating ListMLE Loss (Xia et al., 2008) and Focal Loss (Lin et al., 2017). ListMLE Loss optimizes the global ranking of candidate images, while Focal Loss addresses the challenge of distinguishing between literal and metaphorical meanings by focusing on hard-to-classify samples.

The main works in this paper are as follows:

- **AltCLIP-m18 for Idiomatic Expression Ranking**: We propose AltCLIP-m18 to rank images based on semantic relevance to potential idiomatic expressions in English and Portuguese.
- **LoRA for Efficient Adaptation**: We apply LoRA to AltCLIP-m18, reducing computational costs while maintaining performance.
- **Hybrid Loss for Improved Performance**: By combining ListMLE Loss and Focal Loss, our approach achieves Top-1 Accuracy/DCG scores of 0.53/2.94 for English and 0.77/3.31

*Corresponding author.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2077–2083

for Portuguese on the primary test set, outperforming the base model.

2 Related Works

2.1 Multimodal Alignment Models

Recent advances in multimodal learning have been driven by models like CLIP (Radford et al., 2021), which maps images and text into a shared embedding space through contrastive learning. CLIP's architecture consists of a vision encoder (e.g., ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2021) and a text encoder (Transformer (Vaswani et al., 2017)), enabling effective semantic alignment between visual and textual representations. Extending CLIP to multilingual scenarios, AltCLIP-m18 introduces multilingual contrastive pre-training, supporting 18 languages and achieving state-of-the-art performance in cross-modal tasks. This capability is particularly relevant for SemEval-2025 Task 1 Subtask A, which involves both English and Portuguese, providing a strong baseline for further fine-tuning.

2.2 Parameter-Efficient Fine-Tuning

Fine-tuning large pre-trained models requires significant computational resources. LoRA has emerged as an efficient alternative to address this (Zhang et al., 2024b). LoRA reduces the number of trainable parameters by decomposing the weight update matrix into low-rank components (Hu et al., 2022). This approach allows for efficient adaptation while preserving the model's performance and has been successfully applied in various domains, including natural language processing (Zhang et al., 2024a) and multimodal learning (Shen et al., 2024; Lu et al., 2023).

2.3 Learning-to-Rank Methods

Learning-to-rank (LTR) methods have been extensively studied in information retrieval (Liu et al., 2009), with applications ranging from document ranking to recommendation systems. SemEval-2025 Task 1 Subtask A aims to rank candidate images based on their semantic relevance to a nominal compound (NC) in a given sentence.

Traditional classification or regression losses are ill-suited for this task because they do not directly optimize ranking metrics such as Discounted Cumulative Gain (DCG). Generally, LTR methods can be categorized into the following three paradigms (Liu et al., 2009) :

- **Pointwise Methods**: Treat ranking as a classification or regression problem, focusing on individual samples but ignoring relative order.
- **Pairwise Methods**: Model the relative preferences between pairs of items, capturing local ordering relationships but lacking a global perspective.
- **Listwise Methods**: Optimize the entire ranking list directly, aligning more closely with ranking metrics like DCG.

Among these, Listwise methods, such as ListMLE Loss (Xia et al., 2008), are particularly effective for tasks where global ranking consistency is critical, making them a natural choice for Subtask A.

3 Datasets and Evaluation Metrics

The dataset for SemEval-2025 Task 1 Subtask A includes 70 English and 32 Portuguese training items. Each item contains a context sentence containing a potentially idiomatic NC and five candidate images. The images are categorized into five types: a synonym for the idiomatic meaning, a synonym for the literal meaning, something related to the idiomatic meaning (but not synonymous), something related to the literal meaning (but not synonymous), and a distractor unrelated to both meanings.

For each data item, the primary fields used are **compound** (the idiomatic NC), **sentence_type** (indicating whether the sentence uses the idiomatic or literal sense), **sentence** (the context sentence), **expected_order** (the ground-truth ranking of images), and **image{n}_name** (the filenames of the five candidate images, where n ranges from 1 to 5). In the training data, **sentence_type** and **expected_order** are provided for supervised learning. In the development and test data, these fields are empty, and the model is required to predict **expected_order** based on the context sentence and compound.

The model is evaluated using two key metrics: **Top 1 Accuracy** and **Discounted Cumulative Gain (DCG)**. Top 1 Accuracy measures the model's ability to correctly identify the most representative image for the given context. DCG evaluates the overall ranking quality by assigning higher weights to images ranked closer to the ground-truth top positions. The DCG score is calculated as fol-

lows:

$$DCG = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \quad (1)$$

where rel_i represents the relevance score of the i-th result, and i is the rank position of the result, starting from 1. The term $\log_2(i+1)$ acts as a discount factor, reducing the influence of results that appear later in the ranking.

4 Methodology

Our approach for SemEval-2025 Task 1 Subtask A consists of four key components: (1) the base multimodal model (AltCLIP-m18), (2) parameter-efficient fine-tuning using LoRA, (3) a combined loss function integrating ListMLE Loss and Focal Loss, and (4) a data augmentation strategy to enhance the diversity and robustness of the training data.

4.1 Base Model: AltCLIP-m18

We adopt AltCLIP-m18, a multilingual extension of CLIP, as the base model. AltCLIP-m18 consists of a Transformer-based text encoder and a Vision Transformer (ViT) image encoder, which maps text and images into a shared embedding space. Given a sentence s and an image I , the model computes their similarity score as:

$$\text{sim}(s, I) = \cos(E_{\text{text}}(s), E_{\text{image}}(I)) \quad (2)$$

where E_{text} and E_{image} denote the text and image encoders, respectively, and cos is the cosine similarity function (see Figure 1).

4.2 Parameter-Efficient Fine-Tuning with LoRA

To adapt the pre-trained AltCLIP-m18 model to the task, we employ LoRA, which reduces the number of trainable parameters by decomposing the weight update matrix into low-rank components:

$$\Delta W = A \cdot B \quad (3)$$

where A and B are low-rank matrices with rank r , and ΔW is the weight update. The updated weight matrix is then:

$$W' = W + \alpha \cdot \Delta W \quad (4)$$

where α is a scaling factor that controls the strength of the update.

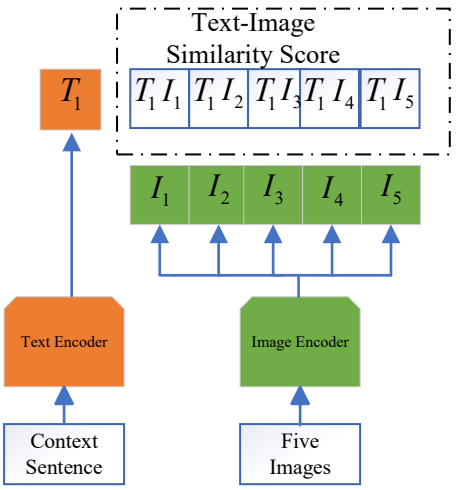

Figure 1: Architecture of AltCLIP-m18 for Text-Image Similarity Computation

In our implementation, LoRA is applied to the query and value projection matrices in the Transformer layers of the text and image encoders (see Figure 2). Research shows that adapting these two projection layers enables effective parameter-efficient tuning (Hu et al., 2022). Detailed hyperparameter configurations are discussed in the Experiments section.

4.3 Combined Loss Function

To optimize the ranking of candidate images, we propose a combined loss function integrating **ListMLE Loss** and **Focal Loss**.

ListMLE Loss maximizes the likelihood of the correct ranking by considering the entire list of candidate images. Given a ground-truth ranking y and a predicted ranking $f(x)$, the loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{ListMLE}} &= -\log P(y|x) \\ &= -\log \prod_{i=1}^{n} \frac{\exp(f(x_{y_i}))}{\sum_{k=i}^{n} \exp(f(x_{y_k}))} \end{aligned} \quad (5)$$

where y_i denotes the i-th item in the ground-truth ranking, $f(x_{y_i})$ is the predicted score for the i-th item, and x is the input to the model.

Focal Loss dynamically adjusts the weight of each sample to emphasize hard-to-classify cases (Lin et al., 2017), i.e., those for which the predicted probabilities are close to 0.5. In our approach, images of NCs with idiomatic and literal interpretations are regarded as hard-to-classify instances.

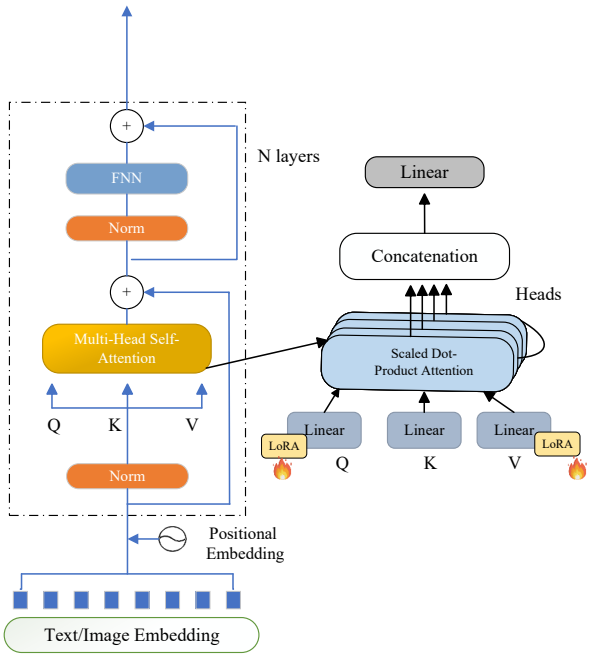

Figure 2: Application of LoRA to Query and Value Projection Matrices.

Misclassification of these cases can significantly impact the Top-1 Accuracy. The loss function is defined as:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{6}$$

where $p_t = p$ for positive samples and $1-p$ for negative samples, with p being the model's confidence in the positive class. γ is the focusing parameter, and α_t is a weighting factor for class balancing. In the expected_order, the first image is treated as the positive sample, while the remaining images are considered negative. This setup allows the model to focus on distinguishing the most representative image from the candidates, thereby improving Top 1 Accuracy.

By combining Focal Loss with ListMLE Loss, our approach optimizes both the overall ranking distribution and the model's ability to handle ambiguous samples. The final loss function is a weighted combination of ListMLE Loss and Focal Loss:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{Focal}} + (1 - \lambda) \cdot \mathcal{L}_{\text{ListMLE}} \tag{7}$$

where λ is a balancing factor.

4.4 Data Augmentation

To enhance the diversity and robustness of the training data, we employ a data augmentation strategy using the DeepSeek-V3 [1] model. For each data point, we generate two sentence variants using carefully designed prompts (see Figure 3): one preserving the original sentence_type and another inverting sentence_type. When sentence_type is inverted, we also invert the top four images in expected_order, ensuring the model learns to distinguish between idiomatic and literal meanings more effectively.

5 Experiments

5.1 Experimental Setup

We trained our model on the augmented dataset with a learning rate of 1×10^{-4}, batch size of 8, and 2 epochs. For Focal Loss, we set $\gamma = 2$ and $\alpha_t = [0.35, 0.1, 0.15, 0.3, 0.1]$ through empirical experiments. For LoRA, we used rank $r = 6$, scaling factor $\alpha = 48$, and dropout rate 0.5 to balance performance and computational efficiency. Table 2 compares the trainable parameters of AltCLIP-m18 between full fine-tuning and the LoRA setup used in our experiments.

5.2 Comparison with Baseline

We compared our approach to the baseline model (AltCLIP-m18) in zero-shot performance. Table 1 presents the baseline results alongside our model's performance with Focal Loss weight $\lambda = 0.15$, while Figure 4 illustrates the impact of different Focal Loss weights on the primary test set.

On the development set, with λ = 0.15, our method achieved identical **Top 1 Accuracy** (0.60) and comparable **DCG** scores to the baseline in both English and Portuguese.

On the primary test set, with λ = 0.15, our model achieved a **Top 1 Accuracy** of 0.53 and **DCG** of 2.94 for English, outperforming the baseline's **Top 1 Accuracy** (0.40) and **DCG** (2.98). For Portuguese, our model achieved a **Top 1 Accuracy** of 0.77 and **DCG** of 3.31, surpassing the baseline's **Top 1 Accuracy** (0.55) and **DCG** (2.98).

On the extended test set, with λ = 0.15, our model achieved a **Top 1 Accuracy** of 0.59 and **DCG** of 2.97 for extended English, outperforming the baseline's **Top 1 Accuracy** (0.57) and **DCG** (2.95), and for extended Portuguese, it matched the baseline's score of 0.53 while maintaining a **DCG** of 2.98.

The improvements over the baseline model stem from Focal Loss and LoRA Fine-Tuning. Focal

[1] https://www.deepseek.com/

For same type variant:

```
type_adverb = "idiomatically" if sentence_type == "idiomatic" else "literally"

prompt = f"Generate a new sentence that includes '{compound}' and is used {type_adverb},
similar to: {sentence}. Provide only the new sentence without any additional text or explanation."
```

For opposite type variant:

```
opposite_adverb = "literally" if sentence_type == "idiomatic" else "idiomatically"

prompt = f"Generate a new sentence that includes '{compound}' but is used {opposite_adverb},
opposite to: {sentence}. Provide only the new sentence without any additional text or explanation."
```

Figure 3: Prompts used for data augmentation.

Table 1: Performance Comparison of Our Approach with Zero-Shot Baseline Across Language Settings

Method	Dev Set				Test Set				Extended Set			
	EN		PT		EN		PT		EN		PT	
	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG
Baseline	0.60	2.89	0.60	3.08	0.40	2.83	0.69	3.22	0.57	2.95	0.53	2.98
Ours($\lambda = 0.15$)	0.60	2.87	0.60	3.00	0.53	2.94	0.77	3.31	0.59	2.97	0.53	2.98

Table 2: Comparison of trainable parameters between full fine-tuning and LoRA for AltCLIP-m18

	Trainable Params	Percentage
Full Fine-tuning	1,194,000,897	100%
LoRA	983,040	0.0823%

Loss improves Top 1 Accuracy by focusing on hard-to-classify samples, while LoRA Fine-Tuning ensures efficient adaptation with minimal computational overhead. Together, they enhance multimodal idiomaticity representation.

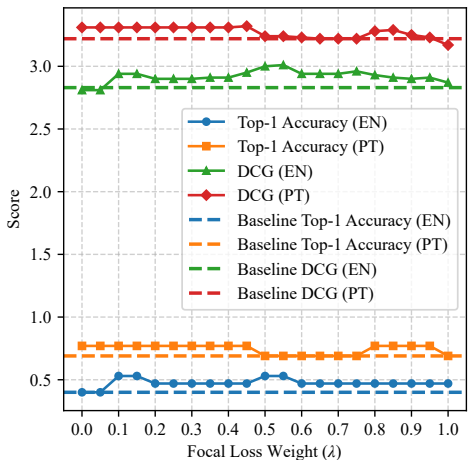

Figure 4: Comparison of Model Performance with Baseline on the Primary Test Set Across Different λ Values

5.3 Ablation Study: Focal Loss Weight λ

We conducted an ablation study to analyze the impact of different Focal Loss weights λ in the combined loss function across development, primary test, and extended test sets. The results are summarized in Table 3.

On the development set, English (EN) showed consistent **Top 1 Accuracy** (0.60) across all λ values, while Portuguese (PT) exhibited more variation, peaking at $\lambda = 0.65$ with **Top 1 Accuracy** **Top 1** (0.77) and **DCG** of 3.13. This stability in development suggests our model's robustness during initial parameter tuning. For the English (EN) primary test set, the best **Top 1 Accuracy** (0.53) was achieved in the vicinity of $\lambda = 0.1$ and $\lambda = 0.5$, while the highest **DCG** (3.01) was observed at $\lambda = 0.55$. In the Portuguese (PT) primary test set, the **Top 1 Accuracy** remained stable at 0.77 for most values of λ, with the **DCG** peaking at 3.32 when $\lambda = 0.45$. For the extended test set, the best **Top 1 Accuracy** (0.59) in the extended English

Table 3: Performance comparison across different λ values. Red highlights indicate the maximum values, while green highlights indicate the minimum values for each metric.

λ	Dev Set				Test Set				Extended Set			
	EN		PT		EN		PT		EN		PT	
	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG	Top 1 Acc	DCG
0.00	0.60	2.87	0.60	3.06	0.40	2.81	0.77	3.31	0.58	2.95	0.55	3.00
0.05	0.60	2.87	0.60	3.05	0.40	2.81	0.77	3.31	0.58	2.96	0.55	3.00
0.10	0.60	2.87	0.60	3.02	0.53	2.94	0.77	3.31	0.59	2.97	0.51	2.97
0.15	0.60	2.87	0.60	3.00	0.53	2.94	0.77	3.31	0.59	2.97	0.53	2.98
0.20	0.60	2.87	0.50	2.98	0.47	2.90	0.77	3.31	0.59	2.96	0.51	2.96
0.25	0.60	2.87	0.50	2.98	0.47	2.90	0.77	3.31	0.59	2.97	0.51	2.96
0.30	0.60	2.88	0.50	2.98	0.47	2.90	0.77	3.31	0.58	2.96	0.51	2.96
0.35	0.60	2.89	0.50	2.98	0.47	2.91	0.77	3.31	0.58	2.96	0.51	2.96
0.40	0.60	2.89	0.50	2.98	0.47	2.91	0.77	3.32	0.57	2.95	0.53	2.98
0.45	0.60	2.90	0.50	2.96	0.47	2.95	0.77	3.32	0.57	2.95	0.53	3.00
0.50	0.60	2.91	0.50	2.97	0.53	3.00	0.69	3.24	0.55	2.93	0.53	2.98
0.55	0.60	2.93	0.50	2.93	0.53	3.01	0.69	3.24	0.55	2.93	0.53	2.98
0.60	0.60	2.92	0.60	3.00	0.47	2.94	0.69	3.23	0.56	2.94	0.55	2.99
0.65	0.60	2.89	0.70	3.13	0.47	2.94	0.69	3.22	0.57	2.95	0.55	2.99
0.70	0.60	2.88	0.70	3.09	0.47	2.94	0.69	3.22	0.57	2.94	0.51	2.99
0.75	0.60	2.88	0.70	3.09	0.47	2.96	0.69	3.22	0.57	2.93	0.53	2.98
0.80	0.60	2.88	0.70	3.09	0.47	2.93	0.77	3.28	0.57	2.93	0.55	2.98
0.85	0.60	2.88	0.70	3.10	0.47	2.91	0.77	3.29	0.57	2.93	0.55	2.98
0.90	0.60	2.86	0.70	3.11	0.47	2.90	0.77	3.25	0.57	2.94	0.56	2.99
0.95	0.60	2.87	0.60	3.03	0.47	2.91	0.77	3.23	0.57	2.94	0.58	3.01
1.00	0.60	2.87	0.60	3.02	0.47	2.87	0.69	3.17	0.57	2.93	0.62	3.04
Average	0.60	2.88	0.59	3.03	0.47	2.92	0.74	3.27	0.57	2.95	0.54	2.98
Std	0.00	0.02	0.08	0.06	0.03	0.05	0.04	0.05	0.01	0.01	0.03	0.02

test set was achieved around $\lambda = 0.1$ and $\lambda = 0.2$. Notably, in the extended Portuguese test set, the highest **Top 1 Accuracy** (0.62) and **DCG** (3.04) were observed at $\lambda = 1$. These results indicate that the optimal value of λ varies across languages and test sets, with a moderate range (e.g., $\lambda = 0.1$ to 0.5) generally balancing ranking performance and classification accuracy.

6 Conclusion

This study proposes a multilingual and parameter-efficient approach for SemEval-2025 Task 1 Subtask A, leveraging AltCLIP-m18, LoRA fine-tuning, and a combined loss function of ListMLE Loss and Focal Loss. The experiments demonstrate significant improvements over the baseline model. However, it is important to acknowledge the limitations of our study. One key limitation is the relatively small size of the training dataset, especially for Portuguese, which may affect the generalizability of our results. Future work could address this by expanding the dataset or exploring transfer learning techniques to leverage larger, related datasets.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

References

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 7139–7159.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. *arXiv preprint arXiv:2211.06679*.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can Transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 3608–3626.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,

Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2):195–208.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, pages 2980–2988.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*.

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*.

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, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe - Advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

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 *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pages 8748–8763. PMLR.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. 2024. Multimodal instruction tuning with conditional mixture of LoRA. *arXiv preprint arXiv:2402.15896*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

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. *Computer Speech & Language*, 19(4):365–377.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pages 1192–1199.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024a. Improving personalized sentiment representation with knowledge-enhanced and parameter-efficient layer normalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8877–8889.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024b. Personalized LoRA for human-centered text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*, pages 19588–19596.