Modgenix at SemEval-2025 Task 1: Context Aware Vision Language Ranking (CAViLR) for Multimodal Idiomaticity Understanding

Joydeb Mondal
AI Researcher
joydeb.mondal@oracle.com

Pramir Sarkar
AI Researcher
parmir.sarkar@oracle.com

Abstract

This paper introduces **Context-Aware Vision Language Ranking (CAViLR)**, a novel multimodal approach designed to tackle the challenges of idiomaticity understanding in both text and images for SemEval-2025 Task 1. In Task 1a, our method ranks a set of images based on their relevance to an idiomatic compound in a given context sentence. For Task 1b, we extend this approach by predicting the final image in a sequence and disambiguating whether the idiom is used figuratively or literally. By leveraging state-of-the-art vision-language models like **CLIP**, **Pixtral-12B**, and **Phi-3.5**, along with a **Mixture of Experts (MoE)** framework, CAViLR effectively integrates multimodal information. Our system demonstrates improved performance in both tasks, offering significant advancements in bridging visual and textual semantics and addressing the complexities of idiomatic expressions.

1 Introduction

Idiomatic Expressions (IEs) present a unique challenge in natural language processing. Their meanings often cannot be deduced from individual words, making them difficult for computational models to process (Mi et al., 2024). Unlike literal expressions, IEs require understanding of linguistic conventions, context, and sometimes cultural nuances (Hajiyeva, 2024). Given their prevalence, accurately interpreting idioms is essential for various NLP tasks such as fact-checking, hate speech detection, sentiment analysis, machine translation, and question-answering (Yosef et al., 2023; Tan and Jiang, 2021). Misinterpreting idiomatic meaning can lead to significant errors, impacting the accuracy and reliability of these applications.

The AdMIRe challenge (Pickard et al., 2025) pushes the boundaries of multimodal understanding by focusing on idiomatic nominal compounds in rich visual contexts. The task presents two main challenges:

- **Task 1a: Image Ranking** — Rank five images based on how well they represent the intended sense of an idiomatic compound in a context sentence.
- **Task 1b: Image Sequence Prediction and Idiom Disambiguation** — Predict the final image in a sequence and determine whether the idiom is used literally or figuratively.

We propose a novel approach, **Context-Aware Vision-Language Ranking (CAViLR)**, that integrates **CLIP** as a baseline model with a **Mixture of Experts (MoE)** framework. This hybrid ensemble method addresses the challenges of understanding idiomatic expressions in visual contexts.

Our approach operates in two stages:

1. **Baseline Model (CLIP):** We first use **CLIP** for both image ranking and sequence prediction. CLIP provides a strong foundation by mapping both text and images into a shared embedding space (Kulkarni et al., 2024).
2. **Hybrid Model with MoE:** We then enhance performance using a **Mixture of Experts (MoE)** framework, where the model dynamically selects expert models like **Pixtral-12B** for visual-textual understanding (Agrawal et al., 2024) and **Phi-3.5** for textual analysis, optimizing performance for each task.

This hybrid approach improves both image ranking (Task 1a) and image sequence prediction with idiomatic disambiguation (Task 1b). The **MoE integration** dynamically selects the best expert based on the input, improving performance across tasks.

In the following sections, we detail the dataset and evaluation setup (§3), describe our methodology (§4), present experimental results (§5), and discuss the implications of our findings. Further details can be found on the official task webpage (SemEval-2025 Task 1, 2025).

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 780–784

2 Related Works

While prior research has focused on idiom detection and interpretation from text, the role of multimodality in idiomaticity understanding remains underexplored. Most related studies address figurative language understanding and disambiguation of mixed visual and textual content. Specifically, visual figurative meaning understanding (Saakyan et al., 2024) aims to assess whether a visual premise entails or contradicts a textual hypothesis. In contrast, visual-word sense disambiguation (Raganato et al., 2023) focuses on selecting, from a set of candidate images, those that match the intended meaning of a target word with limited textual context. The AdMIRe challenge (SemEval-2025 Task 1) extends this idea to idiomatic expressions, where idioms are considered as textual descriptions and images that capture their intended meaning.

Transformer architectures, such as **CLIP** (Kulkarni et al., 2024), and visual LLMs like **LLaVA** (Liu et al., 2023), have shown promising performance on various multimodal tasks (Kulkarni et al., 2024; Vaiani et al., 2023; D'Amico et al., 2023; Napolitano et al., 2024). Moreover, advanced visual LLMs, such as **Qwen2.5-VL** (Bai et al., 2025), **Pixtral-12B** (Agrawal et al., 2024), **Phi-3.5** (Abdin et al., 2024) and **Gemini** (The Gemini Team et al., 2023), have been designed to handle multiple images, further enhancing multimodal understanding.

3 Data

This study uses datasets from SemEval-2025 Task 1, with both English and Portuguese data for Subtasks A and B.

3.1 Subtask A - Static Images

The English training data for Subtask A includes 70 items. Each item consists of a sentence with a potentially idiomatic compound and five candidate images, each with a machine-generated caption. The dataset is organized in a `subtask_a_train.tsv` file containing the following fields:

- `compound`: The potentially idiomatic noun compound.
- `sentence`: The target sentence.
- `image{n}_name`: The filenames of candidate images.
- `image{n}_caption`: Descriptive captions for each image.

Each compound has a corresponding subfolder with 5 images.

The Portuguese training data for Subtask A follows the same structure, with 32 items. Additionally, it includes a `image{n}_caption_pt` column for Portuguese captions.

3.2 Subtask B - Sequences

For Subtask B, the English dataset contains 20 items, each consisting of a sequence of images to complete. The `subtask_b_train.tsv` file includes the following fields:

- `compound`: The idiomatic compound.
- `sequence_caption1`, `sequence_caption2`: Descriptive captions for the first two sequence images.
- `expected_item`: The filename of the image that completes the sequence.

Each subfolder contains 6 images, including two sequence images and 4 candidates.

The Portuguese data for Subtask B is structured similarly to the English dataset.

For further details, refer to the official webpage (SemEval-2025 Task 1, 2025).

4 Methodology

Our approach integrates textual and visual features to rank images and predict image sequences based on their relevance to idiomatic expressions in context. We propose a hybrid method, using a **Mixture of Experts (MoE)** approach, where multiple expert models are selectively used for different tasks based on the input data. The methodology is divided into separate steps for both Subtask A and Subtask B.

4.1 Data Preprocessing

- **Text**: Tokenize sentences using the BERT tokenizer.
- **Images**: Normalize and resize images to 224x224 pixels.

4.2 Feature Extraction

For both Subtask A and Subtask B, textual and visual features are extracted separately:

- **Textual Features**: Extract sentence embeddings using the BERT model (768-dimensional).

- **Visual Features**: Extract image embeddings using **ResNet-50** (2048-dimensional) or **ViT** for CLIP (depending on the model choice).

4.3 Models

After feature extraction, we introduce two main model types: the **Baseline Model** and **Language-Visual Model (LMM)**. Both are utilized in a hybrid fashion to process textual and visual inputs.

4.3.1 Baseline Model: CLIP

For the baseline model, we use **CLIP** (Contrastive Language-Image Pre-training), which is an open-source multimodal model that learns a joint embedding space for images and text. CLIP is pre-trained on a large dataset and can be fine-tuned to perform both image ranking and image sequence prediction tasks effectively.

- **Text Encoder**: CLIP's text encoder is based on a Transformer model (similar to BERT), which converts input sentences into embeddings.
- **Image Encoder**: CLIP uses a **Vision Transformer (ViT)** to generate image embeddings.
- **Contrastive Learning**: CLIP uses contrastive learning to match images with their corresponding textual descriptions.

4.3.2 Language-Visual Model (LMM) with MoE

For more advanced performance, we adopt a hybrid model with the **Mixture of Experts (MoE)** approach. The MoE model dynamically selects expert models based on the task and input context, enabling specialization for different subtasks.

- **Model Selection**: Utilize **Pixtral-12B** for visual-textual understanding, and **Phi-3.5** for textual interpretation. Both models are part of the expert set in the MoE framework.
- **MoE Integration**: The MoE framework selects the most appropriate expert (Pixtral or Phi-3.5) based on the input data, enhancing the performance for each task.
- **Expert Specialization**: **Pixtral** specializes in visual analysis, while **Phi-3.5** is used for more detailed textual analysis.

4.4 Task-Specific Architecture

4.4.1 Subtask A: Image Ranking

For **Subtask A**, the goal is to rank images based on their relevance to the given sentence. After multimodal feature extraction, the model ranks images by processing the combined textual and visual embeddings.

- **Text Encoder**: BERT (or CLIP's text encoder) generates sentence embeddings.
- **Image Encoder**: ResNet-50 (or CLIP's image encoder) generates image embeddings.
- **Fusion Layer**: Combines text and image embeddings into a unified representation.
- **Ranking Layer**: A fully connected neural network that outputs a ranking score for each image.

4.4.2 Subtask B: Image Sequence Prediction

For **Subtask B**, the objective is to predict the next image in a sequence based on the previous images and the sentence context. The **LMM with MoE** approach is applied to both textual and visual embeddings to predict the correct image sequence.

- **Sequence Prediction**: The model is trained to predict the image that logically completes a sequence.
- **Textual and Visual Embeddings**: Embeddings from both modalities are used to predict the sequence of images.
- **MoE Layer**: The MoE model selects the appropriate expert based on the sequence and context.

4.5 Training and Evaluation

For both Subtasks A and B, the models are fine-tuned using a **supervised learning** approach. The training process focuses on minimizing the loss between the predicted ranking or sequence and the ground truth.

- **Loss Function**: For Subtask A, **Mean Squared Error (MSE)** is used to optimize ranking accuracy. For Subtask B, **Cross-Entropy Loss** is used to predict the correct image in the sequence.

- **Optimizer**: We use the **Adam optimizer** with learning rate tuning to optimize the model parameters. The learning rate is dynamically adjusted during training for better convergence.
- **Evaluation Metrics**:
  - **Subtask A**: Evaluation is based on ranking accuracy, with the model's ability to rank images correctly in terms of relevance to the sentence.
  - **Subtask B**: Evaluation is based on sequence prediction accuracy, with the model's ability to predict the next image in the correct order.

5 Experiments and Results

In this section, we present the results of our approach for both Subtask A and Subtask B. We begin by evaluating individual models and progressively combine them into a hybrid approach, which yields the best results.

5.1 Subtask A - Image Ranking

We started with a basic model for ranking images, then gradually built up our hybrid model by integrating Pixtral, Phi-3.5, and baseline components. Below are the results:

Model Type	Score
Baseline Model (Text + Image)	0.33
Pixtral Model	0.47
Phi-3.5 Model	0.47

Table 1: Results for Subtask A - Individual Models

Next, we combined Pixtral, Phi-3.5, and the baseline model to form a hybrid approach:

Hybrid Model Type	Score
Pixtral + Phi-3.5 + Baseline Model	**0.53**

Table 2: Results for Subtask A - Hybrid Model (Pixtral + Phi-3.5 + Baseline)

The hybrid model achieved the best result with a score of 0.53, outpacing the individual models.

5.2 Subtask B - Image Sequence Prediction and Idiom Disambiguation

Similarly, we started with individual models for Subtask B and progressively added components to improve performance. Below are the results:

Model Type	Score
Baseline Model (Text + Image)	0.40
Pixtral Model	0.47
Phi-3.5 Model	0.47

Table 3: Results for Subtask B - Individual Models

Hybrid Model Type	Score
Pixtral + Phi-3.5 + Baseline Model	**0.60**

Table 4: Results for Subtask B - Hybrid Model (Pixtral + Phi-3.5 + Baseline)

Next, we combined Pixtral, Phi-3.5, and the baseline model to form the hybrid approach:

The hybrid model once again achieved the best result, with a score of 0.60.

These results demonstrate that the hybrid approach, integrating Pixtral, Phi-3.5, and the baseline model, provides significant improvements in both image ranking and sequence prediction tasks.

6 Discussion

In **Subtask A** (Image Ranking), the main challenge lies in the model's ability to accurately ground idiomatic expressions in both visual and textual contexts. We observed that:

- Some images strongly match the intended meaning, while others introduce ambiguity.
- Fine-grained semantic differences often lead to misrankings, especially in the case of subtle idiomatic nuances.

For **Subtask B** (Image Sequence Prediction), sequence prediction becomes difficult when idioms are used figuratively or literally. The **MoE framework** helps by dynamically selecting the most suitable expert model for different tasks. However, **CLIP** as the baseline struggles with more complex idiomatic interpretations.

Future work will focus on improving the **MoE** framework, enhancing multimodal embeddings, and developing better disambiguation strategies for figurative and literal meanings.

7 Conclusion

We introduced **CAViLR**, a context-aware, hybrid multimodal approach for image ranking and sequence prediction in **SemEval-2025 Task 1**. By combining **CLIP** with a **Mixture of Experts**

(MoE) framework, our method improves task performance by dynamically selecting expert models. While our approach shows promising results, further refinement is needed to handle subtle idiomatic variations and enhance disambiguation between figurative and literal meanings. Our work contributes valuable insights into developing more robust vision-language models.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Shengyu Bai, Kailin Chen, Xin Liu, Jing Wang, Wenqiang Ge, Shuyang Song, Kaiyu Dang, Pengfei Wang, Shusheng Wang, Jian Tang, Haoyu Zhong, Yuchen Zhu, Jianyu Wan, Penghao Wang, Wen Ding, Zhefu Fu, Yiheng Xu, Jun Ye, Xiaohui Zhang, and 6 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2505.14201*.

Anna D'Amico, Marco Bertoldi, and He Li. 2023. Leveraging vision–language pretraining for improved multimodal understanding. In *Proceedings of the 2023 Conference of the Association for Computational Linguistics*.

Bulbul Hajiyeva. 2024. Challenges in understanding idiomatic expressions. *Acta Globalis Humanitatis et Linguarum*, 1(2):67–73.

Shreyas Kulkarni, Gaurav Chittar, and Robert Sim. 2024. Multimodal language and vision understanding with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hengyuan Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2305.07014*.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context. *arXiv preprint arXiv:2410.16069*.

Giovanni Napolitano, Zixuan Xu, and Xiaodan Chen. 2024. Unifying vision and language for idiomaticity in multimodal tasks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

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)*, Vienna, Austria. Association for Computational Linguistics.

Alessandro Raganato, Iacer Calixto, Atsushi Ushio, José Camacho-Collados, and Mohammad Taher Pilehvar. 2023. Semeval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tanmay Chakrabarty, and Smaranda Muresan. 2024. Understanding figurative meaning through explainable visual entailment. *arXiv preprint arXiv:2401.12072*.

SemEval-2025 Task 1. 2025. Admire: Advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria.

Xiao Tan and Yuan Jiang. 2021. Exploring idioms in question answering systems. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL*.

The Gemini Team and 1 others. 2023. Gemini. *arXiv preprint arXiv:2312.11805*.

Alessandro Vaiani, Ying Zhang, and Kevin Johnson. 2023. Advancing visual language models with multitask learning. In *Proceedings of Neural Information Processing Systems (NeurIPS)*.

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.