FJWU_Squad at SemEval-2025 Task 1: An Idiom Visual Understanding Dataset for Idiom Learning

Maira Khatoon, Arooj Kiyani, Sadaf Abdul Rauf and Tehmina Doltana Farid
Department of Computer Science, Fatima Jinnah Women University, Pakistan
{khatoonmaira629,aroojkiyani12,sadaf.abdulrauf,tehminafarid556}@gmail.com

Abstract

Idiomatic expressions pose a significant challenge in Natural Language Processing (NLP) due to their non-compositional nature which requires contextual understanding beyond literal interpretation. This paper presents our participation in theSemEval-2025 Task 1 on Advancing Multimodal Idiomaticity Representation, where we focused on dataset augmentation and text versus multimodal LLM models. We constructed an enriched idiom-image dataset using human augmented prompt engineering and AI-based image generation models. Performance of textual and vision-language models (VLMs) was compared in ranking images corresponding to idiomatic expressions. Our findings highlight the benefits of incorporating multimodal context for improved idiom comprehension.

1 Introduction

This paper presents *FJWUSemEvalSquad* participation in SemEval-2025 AdMIRe task which focused on improving idiomatic expression understanding in multimodal contexts. Idiomatic expressions are an integral part of natural language which are characterized by their non-compositionality, where the meaning cannot be directly inferred from the individual words (Fazly et al., 2009). For instance, the phrase "spill the beans" does not refer to physically dropping beans but instead conveys the figurative meaning of revealing a secret.

Understanding idioms correctly is essential for various NLP applications including machine translation (Fadaee and Bisazza, 2018; Baziotis et al., 2023; Liu et al., 2023), sentiment analysis (Boag et al., 2015) and conversational AI (Su et al., 2018; Bergman et al., 2022). While large language models (LLMs) such as BERT (Devlin et al., 2019a) and ALBERT (Lan et al., 2020) have improved text-based semantic interpretation, they still struggle with idiomatic expressions due to their reliance on compositionality-based learning (Dankers et al., 2022).

Recent studies in multimodal learning suggest that visual context can significantly enhance the detection of idiomaticity by providing additional semantic cues (Chakrabarty et al., 2022). However, approaches primarily relying on text-based embeddings like BERT and T5 (Devlin et al., 2019b) lack robust mechanisms for leveraging multimodal information effectively. Advancements in large visual language models have focused on improving the alignment between visual and textual modalities (Geigle et al., 2024; Maaz et al., 2024).Visual augmentation leverages vision-language models (VLMs) to associate idioms with relevant imagery, strengthening contextual learning (Tan and Bansal, 2019). Approaches like CLIP and BLIP improve cross-modal alignment, enhancing interpretability.

Contrastive Language-Image Pretraining (ALBEF) Jiang et al. (2023) is one such vision-language model that learns visual-semantic representations by jointly training on large-scale image-text pairs.It employs contrastive learning to align textual descriptions with corresponding images, enabling zero-shot transfer learning across various tasks. CLIP has demonstrated strong performance in understanding abstract and figurative language by associating idioms with relevant imagery, improving multimodal reasoning in NLP applications (Ghosh et al., 2024).

Dataset augmentation is one of the most effective techniques in NLP to improve model performance by enhancing generalization, reducing overfitting, and increasing robustness to variations in language (Sarhan et al., 2022). We aimed to augment the idiom dataset to enhance the ability of models to understand figurative language by expanding the training data using text-based and visual approaches. Our contributions include:

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1759–1765
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

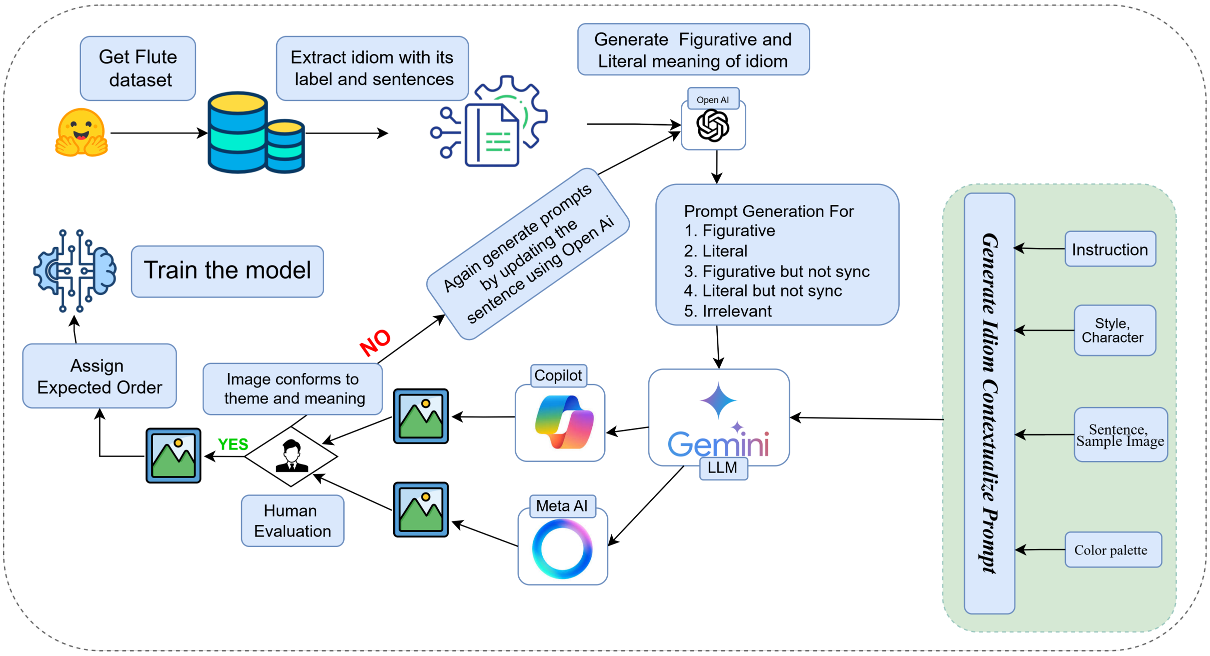

Get Flute dataset
Extract idiom with its label and sentences
Generate Figurative and Literal meaning of idiom
Open AI
Prompt Generation For
1. Figurative
2. Literal
3. Figurative but not sync
4. Literal but not sync
5. Irrelevant
Generate Idiom Contextualize Prompt
Instruction
Style, Character
Sentence, Sample Image
Color palette
Train the model
Again generate prompts by updating the sentence using Open Ai
NO
Assign Expected Order
Image conforms to theme and meaning
YES
Copilot
Gemini
LLM
Meta AI
Human Evaluation

Figure 1: Our approach to systematically generate in theme image prompts and their corresponding idiomatic images leveraging LLMs like OpenAI, Meta AI and human evaluation.

- A visual idiom image dataset[1] shared with research community.
- Comparison of text-based versus vision large language models

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 presents the task description, including subtasks, and Section 4 onwards outlines our dataset and experiments.

2 Related Work

Idiomatic expressions have been a long-standing challenge in NLP due to their semantic opacity and contextual dependency (Sag et al., 2002). Early research relied on frequency-based heuristics and rule-based approaches, which struggled with generalization across diverse idioms (Villavicencio et al., 2005). Advent of deep learning and transformer-based models such as BERT and GPT have demonstrated improvements in idiom classification (Ghosh et al., 2015; Liu et al., 2016). However, these models often fail to distinguish between literal and figurative meanings, especially in context-dependent scenarios (Shwartz and Dagan, 2018).

Text-based models, such as BERT, process sequential data to capture the syntactic and semantics of language (Devlin et al., 2019a). In contrast, image-based models like Vision Transformers (ViTs), analyze visual data by dividing images into patches and processing them to understand spatial relationships. Multimodal learning has emerged as a promising direction to address this limitation by integrating visual and textual representations to enhance NLP models (Kiela et al., 2023). Recent studies have explored vision-language models such as CLIP (Geigle et al., 2024; Maaz et al., 2024) to improve idiomatic language interpretation (Chakrabarty et al., 2022). However, the systematic incorporation of multimodal signals in idiomaticity detection remains an open research problem.

3 Task Description

Subtask A Static Image Ranking: In this subtask, participants are provided with a context sentence containing a potentially idiomatic nominal compound (NC) and five images, each of which could correspond to either the literal or figurative meaning of the expression. The objective was to rank these images on the basis of their relevance to the given idiomatic expression.

[1] https://github.com/sabdul111/Fjwu-Visual-Idioms-SemEval2025

4 Visual Idiom Dataset

The task organizers provided idioms along with their corresponding images for the two subtasks (200 idioms for subtaskA and 70 for subtask B). For each idiom, there were five images representing the meaning in *A:Literal, B:Figurative, C:Literal but Not Synchronized , D:Figurative but Not Synchronized and E:Irrelevant* senses. Task images followed a typical color scheme with brown, yellow and orange in dominance. These idiomatic images used distinctive animated characters.

Figure 1 summarizes our approach. To increase the diversity of the dataset, we systematically selected idiomatic expressions from FLUTE,[2] which is an open repository by ColumbiaNLP. It is a collection of metaphors, similes, sarcasm, idioms, and creative paraphrases. We automatically extracted the relevant fields which included the idiom itself, its associated label, a detailed explanation, a contextual sentence, and its corresponding interpretation. Task idioms were based solely on Nominal Compounds (NC) e.g. *night owl*, whereas we did not make any such distinction and added mutli word idioms too.

4.1 Prompt Tuning using Human Feedback

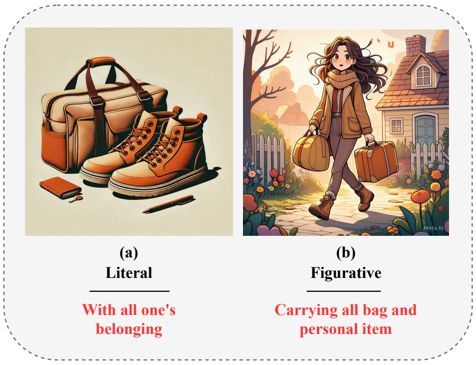
Prompt generation leveraged LLM generation which was verified and tuned by human evaluation to generate the images in line with the task theme as shown in Figures 4 and 2. For each idiom, we generate five images representing different aspects of its meaning as defined in the task. Google Gemini was provided with the sample sentences and reference images to enable visual theme learning. When a reference image is available, Gemini analyzes its artistic style, color palette, and key visual attributes. The extracted style description serves as a guideline for prompt creation. If no reference image is available, a predefined default style is applied.

The ranking of the five generated images was based on human evaluator ratings for relevance, idiom clarity, and visual theme consistency. Figure 4 illustrates the distinction between *literal* and *figurative* meaning. The *literal* meaning corresponds to the image right side (a) with a young woman carrying bags and belongings while leaving a house, directly aligning with the explicit action of "carrying all one's belongings." This is a straightforward representation without any hidden or symbolic interpretation. In contrast, the *figurative* meaning corresponds to the image left side (b), showing various personal items, including shoes, bags, and a notebook, symbolizing the concept of "having all of one's belongings" in a more abstract way than depicting an action. As shown in Figures 2 and 3, human-tweaked prompts significantly enhanced the semantic relevance, stylistic consistency, and idiomatic clarity of the generated images. This demonstrates the importance of human intervention for achieving accurate and task-aligned visual representations.

.

If the *literal but not synchronized* meaning prompt was used, the image might still depict a person carrying items but in an unrelated scenario, such as a delivery worker handling packages, making it misaligned with the idiomatic context. Similarly, a *figurative but not synchronized* meaning prompt might feature an unrelated symbolic image, such as an empty house after someone moves out, conveying the idiom's essence but lacking direct coherence with its intended usage. An *irrelevant* meaning prompt, on the other hand, would fail to relate to either meaning, such as a random travel scene or a landscape with no clear connection to the idiom.

All annotations were manually evaluated by at least two annotators, followed by an adjudication by a third reviewer where necessary.


(a)
Literal
With all one's belonging

(b)
Figurative
Carrying all bag and personal item

Figure 4: Illustration of image generation for literal versus figurative meaning

4.2 Image Generation

Prompt generation was followed by generation using multiple AI models, each contributing unique capabilities to enhance the visual representation of

[2] https://huggingface.co/datasets/ColumbiaNLP/FLUTE

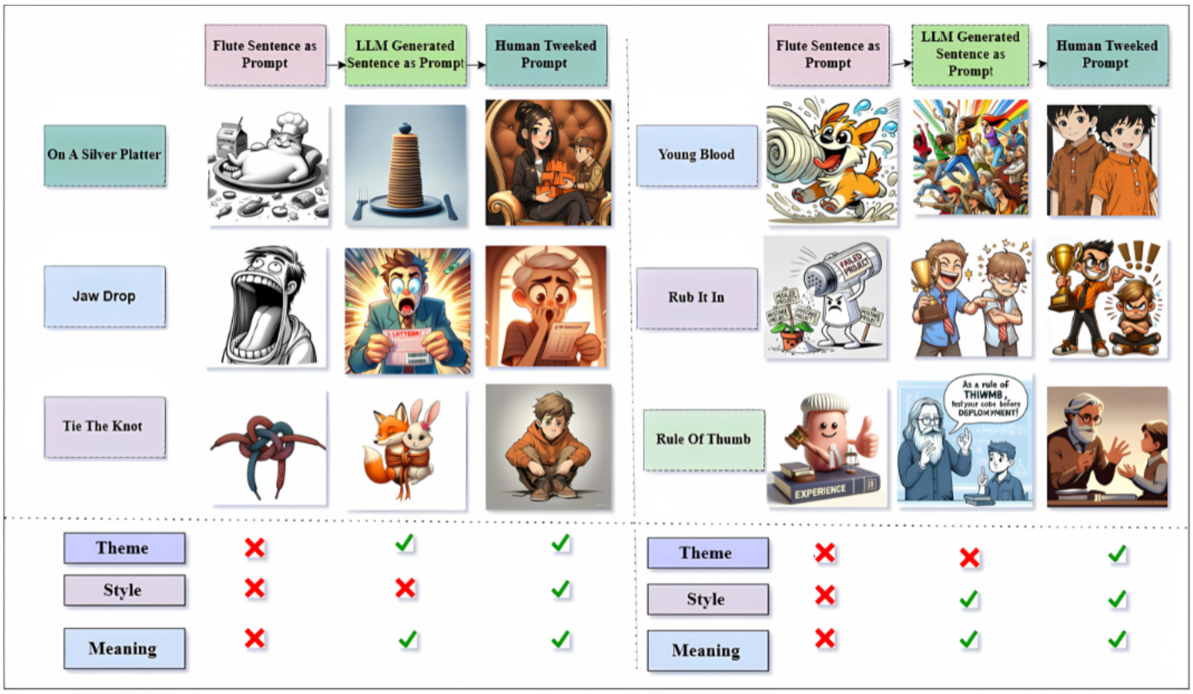


Figure 2: Visual comparison of idiom representations across different prompts

Idiom	Flute sentences as Prompt	LLM generated sentences as Prompt	Human Tweaked Prompt
on a silver platter	Cartoonish style, simple, white background. A lazy cat lounging on a plush silver platter, a fish jumping onto it effortlessly. The fish is perfectly cooked and garnished. Surrounding the platter are scattered half-finished projects – a paintbrush, a book, a puzzle. The cat smirks contentedly. Illustrate the ease with which the cat receives the "reward" (fish) without any effort, contrasting with the surrounding undone tasks, visually representing unearned success or privilege.	Cartoonish illustration, white background: A vibrant, energetic group of diverse young people, radiating enthusiasm, injecting fresh ideas into a tired, grey, older boardroom. They're presenting innovative, colorful charts and graphs, contrasting sharply with the dull surroundings. Focus on the energy exchange; older figures slightly awestruck. Emphasis on playful yet professional demeanor. Avoid literal blood imagery.	A social media influencer girl cartoonic illustration lounges on a luxurious chair, as a delivery worker cartoonic illustration hands them free designer bags. Use just orange, brown, black colour, no noise.

Figure 3: Variations in prompt formulations for generating idiomatic visual representations.

idioms. Meta AI (Meta AI, 2024) was utilized to generate contextually rich images, while Bing AI (Bing AI, 2024) provided diverse visual interpretations of idioms. Microsoft Copilot (Microsoft Copilot, 2024) played a role in assisting with AI-based content refinement.

Figure 2 illustrates the impact of different prompt-generation methods on idiomatic image representations. Initially, captions were generated based on sentences from the FLUTE dataset. However, the resulting images failed to capture the intended meaning, theme, and stylistic coherence of the idioms, as can be seen from first columns in Figure 2. Lower section presents the qualitative analysis by evaluating the generated images against three key criteria: theme, style, and meaning.

To improve representational accuracy, the sentences were refined using OpenAI. The images generated from these LLM-enhanced prompts demonstrated a significant improvement in conveying the idiomatic meaning. However, while the semantic representation improved, the generated images still lacked alignment with the tasks dataset's thematic and stylistic consistency. This highlights the challenge of achieving both semantic fidelity and stylistic coherence in AI-generated idiomatic illustrations. The LLM-generated prompts were then further refined by human evaluation which generated images conveying the intended meaning as well as aligned closely with the tasks thematic and stylistic attributes.

5 Evaluation

We used multiple evaluation metrics to evaluate model performance. *Mean Reciprocal Rank (MRR)* measures how well the model ranks the correct image. It is defined as:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{1}$$

where $|Q|$ is the total number of queries and $rank_i$ is the rank of the first correct result for query i.

Top Accuracy (Top Acc.): Determines whether the top-ranked image correctly represents the idiomatic meaning:

$$\text{Top Accuracy} = \frac{\text{Correct Top Predictions}}{\text{Total Queries}} \tag{2}$$

Spearman Rank Correlation: Evaluates the ranking consistency between predicted rankings and ground truth rankings. It is computed as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3}$$

where d_i is the difference between the two rankings of item i, and n is the number of items ranked.

Discounted Cumulative Gain (DCG): Measures ranking quality by emphasizing the importance of highly relevant items appearing earlier:

$$DCG = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \tag{4}$$

where rel_i is the relevance score of item i and p is the position in the ranking.

Overall Accuracy: Computes the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{5}$$

6 Model Development

Two models were built for comparison.

We used transformer-based paraphrase-multilingual MiniLM-L6-v2 embeddings for *text-based model.* CLIP was used as a multimodal LLM which combines language and image representations in a single joint visual semantic embedding space.

The primary method used to rank the order of images, given the noun compound (NC) and the context sentence, involves a combination of CLIP-based embeddings and a trained ranking model. First, the sentence is passed through the CLIP text encoder to generate a 512-dimensional text embedding, while each candidate image is processed through the CLIP image encoder to obtain corresponding 512-dimensional image embeddings. These embeddings are then expanded to 513 dimensions by adding an extra feature to match the input format expected by the trained ranking model.

Both the text and image embeddings are fed into the ranking model, which predicts a relevance score indicating how well each image matches the given sentence. Separately, the cosine similarity between the original CLIP text and image embeddings is calculated and normalized. The final score for each image is computed by averaging the model-predicted score and the normalized CLIP similarity. The images are then ranked in descending order based on these final combined scores. This hybrid approach ensures that the ranking not only captures basic visual-textual similarity but also leverages the model's ability to distinguish subtle differences in literal and figurative meanings.

Our submission scored 0.60 accuracy in both subtasks A and B. Table 1 shows the results for the two models: MiniLM and MiniLm+CLIP. MiniLm scores 0.27 but after integrating MiniLm with CLIP, it scores 0.60. Because CLIP has strong image-text alignment capabilities, helping the model better associate idiomatic/literal sentences with corresponding images.

Discounted Cumulative Gain(DCG) prioritizes correct images appearing earlier. The increases from 2.54 to 2.94 means that MiniLM + CLIP assigns higher relevance scores to correct images by combining both textual and visual embeddings. Classification Accuracy represents how well the model distinguishes between idiomatic/literal sentences. MiniLM with CLIP improves accuracy score to 0.40 while MiniLM scores 0.10, which means it misclassifies the sentence type

Dataset	Top Accuracy	MRR	DCG Score	Spearman Corr.	Accuracy
Minilm	0.27	0.20	2.54	-1.00	0.10
Minilm + CLIP	0.60	0.34	2.94	0.04	0.40

Table 1: Model scores using the automatic evaluation metrics

7 Conclusion

This study explores the integration of textual and visual modalities to improve idiomatic expression understanding within the AdMIRe task. We constructed a visual idiom dataset, incorporating human augmented prompt engineering and AI based image generation. Our experiments highlight the strengths and limitations of both text-based and vision-language models in idiomaticity detection.

References

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700. Association for Computational Linguistics.

A. Stevie Bergman, Gavin Abercrombie, Shannon Spruit, Dirk Hovy, Emily Dinan, Y-Lan Boureau, and Verena Rieser. 2022. Guiding the release of safer e2e conversational ai through value sensitive design. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 29–39. Association for Computational Linguistics.

Bing AI. 2024. Bing ai image creator. Available at: `https://www.bing.com/create`.

William Boag, Peter Potash, and Anna Rumshisky. 2015. TwitterHawk: A feature bucket based approach to sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 640–646. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2022. Multimodal idiomaticity detection using vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4171–4186. Association for Computational Linguistics.

Marzieh Fadaee and Arianna Bisazza. 2018. Examining the role of multiword expressions in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2554–2559. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Lexical cohesion and the identification of non-compositional multiword expressions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 647–655. Association for Computational Linguistics.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2024. mblip: Efficient bootstrapping of multilingual vision-llms. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*.

Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty, and Emmy Liu, editors. 2024. *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*. Association for Computational Linguistics, Mexico City, Mexico (Hybrid).

Swagata Ghosh, Tony Veale, and Andy Way. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*.

Chaoya Jiang, Wei Ye, Haiyang Xu, Songfang Huang, Fei Huang, and Shikun Zhang. 2023. Vision language pre-training by contrastive learning with cross-modal similarity regulation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14660–14679, Toronto, Canada. Association for Computational Linguistics.

Douwe Kiela et al. 2023. The dawn of foundation models for multimodal tasks: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13338–13369, Toronto, Canada. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111. Association for Computational Linguistics.

Qian Liu, Zhen Li, Yuexiang Lin, Chun Yuan, and Xu Sun. 2016. Neural multiword expression identification with bert and character-level features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Meta AI. 2024. Meta ai image generation models. Available at: `https://ai.meta.com/`.

Microsoft Copilot. 2024. Microsoft copilot for content generation. Available at: `https://copilot.microsoft.com/`.

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

Injy Sarhan, Pablo Mosteiro, and Marco Spruit. 2022. Uu-tax at semeval-2022 task 3: Improving the generalizability of language models for taxonomy classification through data augmentation. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, page 271–281. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2018. A sequential neural model for multiword expression identification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Pei-Hao Su, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Stefan Ultes, Tsung-Hsien Wen, and Steve Young. 2018. Deep learning for conversational ai. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 1–5. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Multiword expressions: Challenges and contributions for lexical semantics. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*.