dutir914 at SemEval-2025 Task 1: An integrated approach for Multimodal Idiomaticity Representations

Yanan Wang, Dailin Li, Yicen Tian, Bo Zhang
Jian Wang*, Liang Yang
School of Computer Science and Technology, Dalian University of Technology, China
{wangyanan,ldlbest,yicentian,zhangbo1998}@mail.dlut.edu.cn
{wangjian,liang}@dlut.edu.cn

Abstract

SemEval-2025 Task 1 introduces multimodal datasets for idiomatic expression representation. Subtask A focuses on ranking images based on potentially idiomatic noun compounds in given sentences. Idiom comprehension demands the fusion of visual and auditory elements with contextual semantics, yet existing datasets exhibit phrase-image discordance and culture-specific opacity, impeding cross-modal semantic alignment. To address these challenges, we propose an integrated approach that combines data augmentation and model fine-tuning in subtask A. First, we construct two idiom datasets by generating visual metaphors for idiomatic expressions to fine-tune the CLIP model. Next, We propose a three-stage multimodal chain-of-thought method, fine-tuning Qwen2.5-VL-7B-Instruct to generate rationales and perform inference, alongside zero-shot experiments with Qwen2.5-VL-72B-Instruct. Finally, we integrate the output of different models through a voting mechanism to enhance the accuracy of multimodal semantic matching. This approach achieves **0.92** accuracy on the Portuguese test set and **0.93** on the English test set, ranking **2nd** and **2nd**, respectively. The implementation code is publicly available here[1].

1 Introduction

Idioms, as fixed expressions, are typically understood through multisensory experiences and contextual awareness of the real world, rather than by directly inferring the meaning of individual words. While multimodal learning has emerged as a critical research direction to address this limitation, current models still face challenges in reconciling literal and figurative meanings of idioms (Yosef et al., 2023).

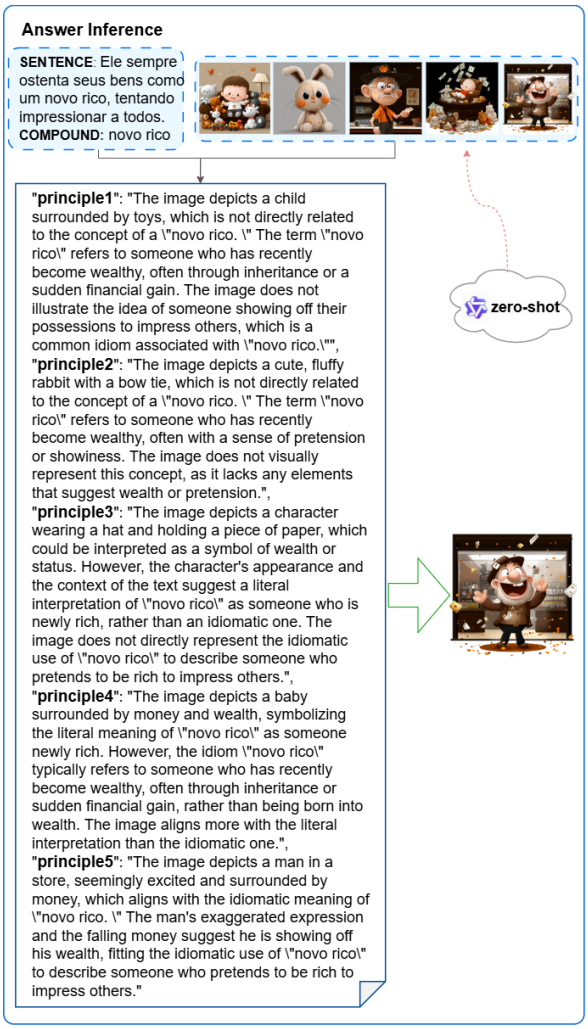

Figure 1: An example of CoT outperforming Zero-Shot Inference in selecting the top image.

Against this backdrop, SemEval-2025 Task 1 : AdMIRe (Advancing Multimodal Idiomaticity Representation) (Pickard et al., 2025) introduces a systematic evaluation framework for multimodal idiomaticity representation. This task builds on the foundation laid by SemEval-2022 Task 2 (Madabushi et al., 2022) in exploring figurative language processing and further advances visual-textual analysis of idiomatic compounds in English and Portuguese.

*Corresponding author.
[1] https://github.com/wyn1015/semeval

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1198–1203

However, the scarcity of cross-lingual multimodal training data for idiomatic compounds hinders model generalization across languages and modalities. To address this, we explore the process of multimodal idiom generation. Specifically, we generate textual explanations and images for compound idiomatic expressions using DeepSeek-V3 (Liu et al., 2024) and Flux.1-dev[2], respectively, constructing two multimodal idiom datasets. These datasets are combined with the original training set to fine-tune the CLIP model (Radford et al., 2021). This enhances its cross-language understanding, allowing it to better capture semantic relationships in multimodal idiomatic expressions.

To further mitigate overfitting risks and improve generalization, we experiment with integrating multimodal large language models. By applying the chain-of-thought principle to Qwen2.5-VL-7B-Instruct (Bai et al., 2025), we generate a step-by-step reasoning method. We integrate textual and visual information into a multi-stage framework that separates the processes of basic principle generation, fine-tuning, and answer inference (see Figure 1). Moreover, we explore the use of zero-shot inference with Qwen2.5-VL-72B-Instruct (Bai et al., 2025).

To leverage these method-specific advantages, we finally design an ensemble approach that combines the outputs of fine-tuned and zero-shot inference models through majority voting, achieving our best results.

2 Background

Recent advances in large language models, such as chain-of-thought prompting (Wei et al., 2022) and zero-shot reasoning (Kojima et al., 2022), have enhanced complex task-solving capabilities. Multimodal reasoning techniques offer new pathways for semantic disambiguation through world knowledge integration. Schwenk et al. (2022) and Zhang et al. (2023) demonstrate that cross-modal approaches surpass single-modality performance. However, the non-compositional nature of idiomatic semantics limits explicit decomposition (Phelps et al., 2024), while dataset artifacts constrain generalization (Boisson et al., 2023). Cultural adaptability solutions employ multilingual prompts (Mu et al., 2025), photorealistic diffusion (Saharia et al., 2022), and reinforcement learning (Xu et al., 2023), though metrics like CLIPScore (Hessel et al., 2021) lack semantic depth. Visual metaphor generation requires dual semantic understanding and cross-modal alignment (Chakrabarty et al., 2023; Yosef et al., 2023; Akula et al., 2023), yet conceptual-imagery inconsistencies persist.

To bridge these gaps, our work introduces multimodal data synthesis and model-scale-aware integration. By generating cross-lingual textual explanations and images, we augment training data for CLIP (Radford et al., 2021), mitigating data scarcity and enhancing cross-language alignment in idiomatic expressions. Further, we integrate multimodal large language models (Qwen2.5-VL-7B/72B (Bai et al., 2025)) through a three-stage chain-of-thought framework. By ensembling the outputs of these models via majority voting, our approach addresses prior limitations in data diversity, cultural adaptability, and model generalization, while systematically exploiting scale-dependent capabilities.

3 System Overview

As depicted in Figure 2, this paper presents a multimodal model ensemble method. First, data augmentation techniques are employed alongside binary classification fine-tuning of the CLIP model. Next, the chain-of-thought strategy is applied for staged fine-tuning and inference of the Qwen2.5-VL-7B-Instruct model while incorporating a zero-shot inference mechanism. Finally, a majority voting strategy from ensemble learning is utilized to combine predictions from multiple models, reducing the bias of individual models, and thus improving the overall system's robustness and accuracy.

3.1 Data Augmentation

Our work is inspired by visual metaphor generation techniques, specifically the combination of large language models with generative models to create visual representations of abstract concepts. This approach drives our exploration of idiomatic expressions, enhancing the understanding of compound word meanings through multimodal integration. We use the DeepSeek-V3 API to generate idiomatic explanations for compound words in English and Portuguese from the SemEval-2022 Task 2 datasets (Madabushi et al., 2022), along with five sentences representing their idiomatic meanings. These sentences are then processed by the Flux.1-dev model to generate images.

[2] https://www.modelscope.cn/models/black-forest-labs/FLUX.1-dev

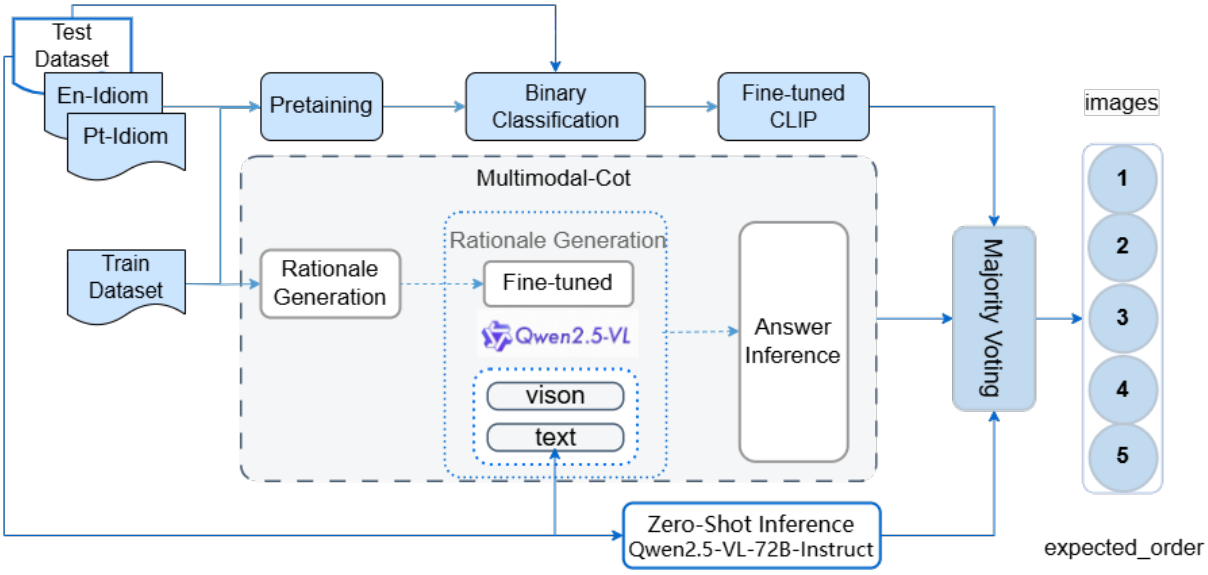

Figure 2: The overall architecture of our ensemble approach.

By pairing the idiomatic explanations with images, we create two multimodal datasets: En-Idiom, containing 243 English compound words and 1,215 images, and Pt-Idiom, consisting of 143 Portuguese compound words and 715 images. After filtering both datasets using CLIP to select image-sentence pairs with a similarity greater than 20, we obtain the Check-Idiom dataset with 1,780 data points. The similarity calculation formula is as follows:

$$s(I,T) = \frac{f_\theta(I)}{\|f_\theta(I)\|} \cdot \frac{g_\phi(T)}{\|g_\phi(T)\|} \quad (1)$$

The input image I and input text T are processed by the image encoder f_θ (using ViT-L/14-336) and the text encoder g_ϕ, respectively.

3.2 CLIP based on Binary classification

The introduction of the CLIP model has provided new perspectives for multimodal reasoning. We propose a binary classification-based fine-tuning method for CLIP, where compound usage in sentences is classified as either literal or idiomatic. This method combines binary classification with image-text pair construction and similarity sorting for more precise multimodal alignment. The steps are as follows:

Image-Text Pair Construction We augment the English and Portuguese training sets with two previously created datasets. If a compound is used idiomatically, its idiomatic meaning is paired with the correct image; if used literally, the sentence is paired with the corresponding image. This ensures that the image-text pairs accurately reflect the different semantic usage scenarios of compounds.

Binary Classification Training During fine-tuning, we select text representations based on compound usage type dynamically.

Similarity Sorting We use Qwen2.5-7B-Instruct to classify compound usage in the test set and generate idiomatic meanings. The fine-tuned CLIP model then ranks images based on the similarity to the corresponding text, considering the usage type.

3.3 Multimodal-CoT

We propose a multimodal chain-of-thought method for rationale generation to enhance the reasoning capabilities of large language models in image-text matching. The method consists of three stages:

Stage 1: Basic Rationale Generation We use Qwen2.5-VL-7B-Instruct with zero-shot chain-of-thought prompting to generate rationales for the correct answers in the English and Portuguese training sets. The model receives prompts with sentences, compounds, usage types, and questions to guide logical reasoning.

Stage 2: Fine-tuning and Rationale Generation The model is fine-tuned on a multimodal dataset of images, prompts, and rationales, enabling it to generate five rationales for each set of five images in the test set.

Stage 3: Answer Inference For each group of 5 images and their corresponding rationales and

prompts without usage types, the final inference is performed to derive the answer based on accumulated multimodal information.

Additionally, we conduct experiments on two-stage zero-shot inference on the test sets, where basic rationales are generated in the first stage, and final inference is performed in the second stage without fine-tuning. We also test zero-shot inference and two-stage zero-shot inference on LLMs with varying parameter sizes.

3.4 Ensemble

The ensemble approach consists of the following parts: (1) The fine-tuned CLIP model. (2) Multi-Modal CoT$_{72B}$: The Qwen2.5-VL-7B-Instruct model is fine-tuned in the first two stages, with the third stage utilizing the API of Qwen2.5-VL-72B-Instruct. (3) Zero-Shot Inference performed by accessing the API of Qwen2.5-VL-72B-Instruct. Specifically, for each position in the expected sequential results of these three methods on the test set and the extended evaluation set, a majority voting mechanism is applied to determine the final images ranking.

4 Experimental Setup

During CLIP training, the maximum text length was set to 77. The batch size was 16, with an initial learning rate of 5e-5 and a warm-up ratio of 0.1. The experiment ran on an RTX 4090 GPU. The CLIP model, fine-tuned on the augmented En-Idiom dataset, performed well on the English test set with 2 epochs. The Check-Idiom dataset, combined with the training set, was used for 3 epochs, achieving good results on the Portuguese test set.

During the multimodal chain-of-thought rationale generation and fine-tuning phase, Qwen2.5-VL-7B-Instruct was fine-tuned on the LLaMA-Factory platform. To ensure rationale accuracy, 127 data points from the English and Portuguese training and development sets were used. The setup involved LoRA fine-tuning with 4-bit quantization and bf16 mixed-precision training. The batch size was 2, initial learning rate 5e-5, temperature 1, top-p 0.01, and max sequence length 10240. During zero-shot inference and two-stage zero-shot inference, either local deployment of Qwen2.5-VL-7B-Instruct or the API of Qwen2.5-VL-72B-Instruct was used, with temperature set to 1 and top-p set to 1. The experiment was conducted on an A40 GPU.

5 Results

In subtask A, the task organizer creates two evaluation metrics.

- **Top Image Accuracy**: Correct identification of the most representative image. The metric presented on the leaderboard is Top 1 Accuracy.
- **Rank Correlation**: Spearman's rank correlation of model rankings with ground truth. However, the metric presented on the leaderboard is DCG Score.

$$\text{DCG} = \sum_{i=1}^{n} \frac{\text{rel}_i}{\log_2(i+1)} \tag{2}$$

where DCG (Discounted Cumulative Gain) measures ranking quality. rel_i is the relevance score of the image at position i, and n is the total number of ranked images. The denominator, $\log_2(i+1)$, serves as a discount factor, reducing the contribution of lower-ranked images to the overall score.

Our ensemble approach reaches 0.93 and 0.92 accuracy in the English and Portuguese test sets, respectively, with DCG scores of 3.46 and 3.43. It also achieves 0.79 and 0.69 accuracy on the extended evaluation sets, confirming the effectiveness of majority voting in integrating the strengths of the fine-tuned CLIP model with the generalization power of multimodal large language models.

Table 1 illustrates the performance of CLIP across different image resolutions and datasets. The results indicate that CLIP's zero-shot reasoning has limited capability for multimodal alignment of idiomatic compounds. After fine-tuning CLIP with data augmentation and binary classification, the test set accuracy is significantly improved, confirming the effectiveness of dynamically adjusting text-matching based on the usage type of compounds. However, considering both languages, the accuracy and DCG scores on the extended set are not high, likely due to the fine-tuning data being biased towards idiomatic types, which restricts generalization.

Table 2 presents the results of our different methods on Portuguese and English datasets. Compared to other methods using the 7B model, Multi-Modal CoT$_{7B}$ achieves the best accuracy and DCG Score on the Portuguese test set and extended evaluation set. This indicates that stage-wise fine-tuning is effective, although smaller models do not experience significant gain.

Model	Training Data	Accuracy(EN)	Accuracy(XE)	Accuracy(PT)	Accuracy(XP)
$CLIP_{224}$	-	0.47	0.46	0.62	0.44
$CLIP_{224}$	En-Idiom+train	0.67	0.42	0.62	0.38
$CLIP_{224}$	Check-Idiom+train	0.47	0.41	0.60	0.40
$CLIP_{336}$	-	0.47	0.46	0.69	0.40
$CLIP_{336}$	En-Idiom+train	0.93	0.46	0.54	0.40
$CLIP_{336}$	Check-Idiom+train	0.73	0.47	0.85	0.36

Table 1: Results of CLIP models on test and extended evaluation sets for different image resolutions and training data (where $CLIP_{224}$ refers to CLIP-ViT-L/14, and $CLIP_{336}$ refers to CLIP-ViT-L/14-336).

Method	Language	Accuracy	DCG Score	Accuracy (XE)	DCG Score (XP)
Zero-$Shot_{7B}$	PT	0.69	3.02	0.53	2.84
Two-$Stage_{7B}$	PT	0.62	2.91	0.51	2.80
Multi-Modal CoT_{7B}	PT	0.85	3.24	0.55	2.89
Zero-$Shot_{72B}$	PT	0.85	3.23	0.71	3.10
Two-$Stage_{72B}$	PT	0.85	3.26	0.49	2.74
Multi-Modal CoT_{72B}	PT	0.85	3.24	0.65	3.02
Ensemble	PT	**0.92**	**3.43**	**0.69**	**3.06**
Zero-$Shot_{7B}$	EN	0.53	2.80	0.56	2.84
Two-$Stage_{7B}$	EN	0.67	2.93	0.53	2.82
Multi-Modal CoT_{7B}	EN	0.53	2.77	0.55	2.87
Zero-$Shot_{72B}$	EN	0.80	3.33	0.80	3.22
Two-$Stage_{72B}$	EN	0.47	2.77	0.64	2.99
Multi-Modal CoT_{72B}	EN	0.60	2.89	0.73	3.15
Ensemble	EN	**0.93**	**3.46**	**0.79**	**3.28**

Table 2: Performance of different methods on the Portuguese and English test and extended evaluation sets.

The 72B model outperforms the 7B model on the extended evaluation set, demonstrating that the model scale has a significant impact on performance. For the 72B models, the Multimodal CoT method, which incorporates multimodal information and performs chain-of-thought reasoning with fine-tuned rationale generation, outperforms the Two-Stage method, thereby enhancing the model's reasoning capabilities. However, Zero-$Shot_{72B}$ performs well on the test set and achieves the best accuracy and DCG Scores on the extended set, validating the zero-shot generalization advantages of large-scale models.

6 Conclusion

This paper proposes an ensemble approach that integrates fine-tuned CLIP, multi-stage chain-of-thought reasoning, and zero-shot inference from large language models, focusing on enhancing the semantic and visual understanding of idiomatic nominal compounds through a multimodal integration framework. Experiments validate the effectiveness of data augmentation for fine-tuning CLIP and highlight the strong generalization capabilities of multimodal large language models. While stage-wise fine-tuning can improve performance compared to two-stage reasoning frameworks, it may still underperform relative to zero-shot inference models. Our future work will be dedicated to optimizing data generation and balancing strategies to mitigate distribution biases, further exploring the interactions between model scale and reasoning stages, and optimizing multimodal semantic understanding of idiomatic expressions.

References

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie

Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. *arXiv preprint arXiv:2311.00790*.

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. *arXiv preprint arXiv:2305.14724*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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*.

Yongyu Mu, Hengyu Li, Junxin Wang, Xiaoxuan Zhou, Chenglong Wang, Yingfeng Luo, Qiaozhi He, Tong Xiao, Guocheng Chen, and Jingbo Zhu. 2025. Boosting text-to-image generation via multilingual prompting in large multimodal models. *arXiv preprint arXiv:2501.07086*.

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. Semeval-2025 task 1: Admire - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, Proceedings of Machine Learning Research. 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 *International conference on machine learning*, pages 8748–8763. PmLR.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. *arXiv preprint arXiv:2303.15445*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.