tiberiucarp at MWE-2026 AdMIRe 2: GLIMMER-Gloss-based Image Multiword Meaning Expression Ranker

Andrei Tiberiu Carp
Tomorrow University
ING Hubs Romania
tiberiucarp@gmail.com

Abstract

Multiword expressions (MWEs), particularly idioms, pose persistent challenges for vision-language systems due to their non-compositional semantics and culturally grounded meanings. This paper presents GLIMMER, a three-stage hybrid ranking system that evaluates how well images express the intended meaning of MWEs across 15 languages. Our approach uses LLM-generated semantic glosses as multilingual meaning anchors, combined with dual-path embedding scoring (textual captions and visual features), and LLM-based semantic verification. Evaluated on the ADMIRE shared task benchmark, GLIMMER achieves competitive performance across diverse languages without relying on parallel training data or language-specific resources. The results show that using glosses to anchor meaning helps match idioms with images across languages and modalities, and that combining retrieval with reasoning is more robust than using embeddings alone.

1 Introduction

Multiword expressions (MWEs), such as idioms, convey meanings that cannot be derived compositionally from their constituent words. Although large language models (LLMs) have shown improved handling of idiomaticity in text-only settings (Tayyar Madabushi et al., 2022; Tedeschi et al., 2022; Tian et al., 2023), multimodal understanding of idioms remains an open challenge. Vision-language models (VLMs) excel in literal and compositional grounding, but often fail when figurative meanings diverge from surface-level visual cues (Yuksekgonul et al., 2022; Akula et al., 2023).

This work, introduces GLIMMER (GLoss-based Image Multiword Meaning Expression Ranker), a hybrid system developed for the ADMIRE shared task (Arslan et al., 2026), which focuses on ranking images according to how well they express the intended (idiomatic or literal) meaning of an MWE in context. GLIMMER is designed around the observation that *idioms are meaning-level units*, and that explicit semantic representations can serve as stable anchors across modalities and languages. Our key contributions are:

1. **Gloss-based semantic anchoring** using LLM-generated contextual definitions as multilingual meaning pivots

2. **Hybrid retrieval-reasoning architecture** combining embedding-based similarity with LLM-based semantic verification

3. **Dual-modality scoring** using both image captions and raw visual features

GLIMMER was evaluated on the ADMIRE shared task benchmark (Torunoğlu-Selamet et al., 2026) and performs competitively in 15 typologically diverse languages without requiring parallel data or task-specific training, highlighting the effectiveness of gloss-centered multimodal reasoning.

The paper is structured as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 presents the experimental setup; Section 5 reports the results and analysis; and Sections 6 and 7 discuss the limitations and conclusions, respectively.

2 Related Work

The detection of multilingual idiomaticity has emerged as a key challenge in NLP (Tayyar Madabushi et al., 2022; Tedeschi et al., 2022). While Transformer models encode idiomatic meanings differently from literal phrases (Tian et al., 2023), recent evaluations show that even LLMs struggle without explicit semantic cues (Phelps et al., 2024). In the visual domain, early vision–language

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 170–176

models (Li et al., 2023a; Huang et al., 2023) excel at compositional tasks but often fail on figurative grounding (Yuksekgonul et al., 2022; Akula et al., 2023; Saakyan et al., 2025). However, recent work demonstrates that textual explanations can act as semantic bridges for non-literal matching (Chakrabarty et al., 2023), motivating our gloss-based design.

Our architecture adapts retrieval-augmented generation (Borgeaud et al., 2022; Izacard et al., 2023) to the multimodal idiom domain. Hybrid pipelines combining dense retrieval with neural reasoning have proven effective for complex semantics (Ni et al., 2025; Mao et al., 2021). To enable zero-shot transfer, we leverage advances in multilingual sentence embeddings (Muennighoff et al., 2023; Li et al., 2023b; Duquenne, 2024). Finally, our approach aligns with findings in multimodal chain-of-thought reasoning (Achiam et al., 2023; Zhang et al., 2023), utilizing LLM-generated glosses as explicit semantic anchors to resolve ambiguity.

3 Methodology

3.1 Problem Formulation

Given a multiword expression e, context sentence s indicating usage type $t \in \{\text{idiomatic}, \text{literal}\}$, and a set of candidate images $\mathcal{I} = \{(I_1, c_1), \ldots, (I_n, c_n)\}$ where I_i is an image and c_i its caption, our goal is to rank images by how well they express the intended meaning of e in context s.

3.2 Three-Stage Pipeline

Stage 1: Gloss Generation We generate a contextual semantic gloss for each MWE using an instruction-tuned LLM (OpenAI GPT-5.1):

> *Given the expression "{e}" used in: "{s}"*
> *Is this idiomatic or literal usage?*
> *Provide a concise gloss explaining the meaning.*

The gloss g serves as a language-independent semantic anchor, cached for efficiency. For unlabeled test data, we infer usage type t via prompting. Despite potential generation noise, g provides a transparent intermediate representation that enhances downstream alignment.

Stage 2: Dual Embedding Scoring For each candidate (I_i, c_i), we compute two complementary similarity scores:

Text Path: Using multilingual sentence transformer (`paraphrase-multilingual-mpnet-base-v2`) (Reimers and Gurevych, 2020):

$$sim_{\text{text}}(i) = \cos(\text{embed}(c_i), \text{embed}(g)) \quad (1)$$

Vision Path: Using CLIP ViT-B-32 (Radford et al., 2021):

$$sim_{\text{clip}}(i) = \cos(\text{CLIP}_{\text{img}}(I_i), \text{CLIP}_{\text{text}}(g)) \quad (2)$$

Combined embedding score:

$$score_{\text{embed}}(i) = 0.6 \cdot sim_{\text{text}}(i) + 0.4 \cdot sim_{\text{clip}}(i) \quad (3)$$

Stage 3: LLM Semantic Verification Embedding-based similarity alone may conflate literal and idiomatic interpretations. We therefore use an LLM (OpenAI GPT-5.1) to perform fine-grained semantic verification, scoring whether a caption describes an image that expresses the gloss meaning:

> *Gloss: "{g}"*
> *Caption: "{c_i}"*
> *Does this caption match the gloss meaning?*
> *Rate 0-100.*

This yields $score_{\text{llm}}(i) \in [0, 100]$. Although this step relies on a proprietary LLM, the prompts are deterministic, and the gloss representations reusable, mitigating variability. We normalize and fuse the scores:

$$score_{\text{final}}(i) = 0.4 \cdot score_{\text{embed}}(i) + 0.6 \cdot \frac{score_{\text{llm}}(i)}{100} \quad (4)$$

The final ranking orders the images by descending $score_{\text{final}}$.

Figure 1 presents how a semantic gloss is generated to anchor the meaning (Stage 1), followed by dual-path embedding scoring (Stage 2) and fine-grained LLM verification (Stage 3).

3.3 Design Rationale

Why glosses? Glosses externalize meaning to enable zero-shot cross-lingual transfer without parallel data. They provide explicit semantic context for evaluation, serving as anchors for non-literal matching as validated in recent visual metaphor research (Chakrabarty et al., 2023).

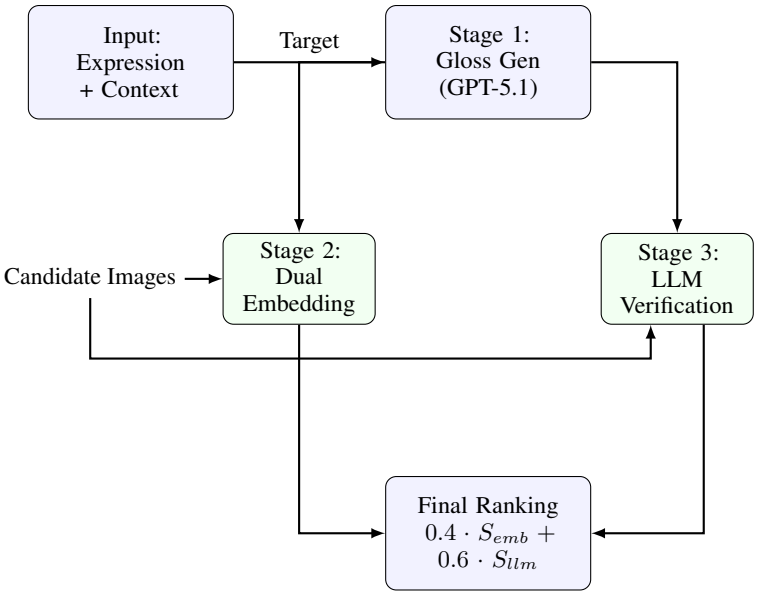

Figure 1: The GLIMMER architecture.

Why hybrid scoring? Fusion combines efficient embedding-based retrieval with precise LLM verification to balance scalability and reasoning depth (Ni et al., 2025). This approach mitigates the tendency of embeddings to conflate literal and idiomatic meanings while avoiding the computational costs of pure LLM scoring.

Why dual modality? Combining captions and images improves ranking robustness. This approach compensates for captions that omit visual cues and raw images that lack linguistic grounding, yielding more reliable retrieval.

Weight tuning We set Text/CLIP weights (60/40) to prioritize captions for abstract semantic clarity. Conversely, Embed/LLM weights (40/60) favor LLM verification to leverage nuanced reasoning for figurative language. These parameters were empirically validated via grid search on development data.

4 Experimental Setup

4.1 Dataset

We evaluate on the ADMIRE shared task dataset (Pickard et al., 2025), covering 15 languages: Chinese (ZH), Georgian (KA), Greek (EL), Igbo (IG), Kazakh (KK), Norwegian (NO), Portuguese-Brazil (PT-BR), Portuguese-Portugal (PT-PT), Russian (RU), Serbian (SR), Slovak (SK), Slovenian (SL), Spanish-Ecuador (ES-EC), Turkish (TR), and Uzbek (UZ).

For each language, the dataset provides:

- Multiword expressions with context sentences
- Sets of 5 candidate images per expression
- Image captions in the target language
- Usage type labels (idiomatic/literal) for training only

The test set contains expressions without usage type labels, requiring automatic inference. The evaluation metrics used are: Accuracy (Acc), Spearman Correlation, and Normalized Discounted Cumulative Gain (nDCG).

Metrics are computed per usage type (idiomatic/literal) and aggregated across languages.

4.2 Implementation

Our system is implemented in Python using SentenceTransformers and OpenAI libraries, with the following configuration:

- **LLM**: OpenAI GPT-5.1 via Responses API (temperature 0.7 for glosses, 0.0 for scoring)
- **Sentence Encoder**: paraphrase-multilingual-mpnet-base-v2
- **CLIP**: OpenAI ViT-B-32 with default preprocessing
- **Weight Parameters**: Text/CLIP $\alpha = 0.6$, Embedding/LLM $\alpha = 0.4$
- **Gloss Caching**: Enabled to reduce API calls (same expression $\rightarrow$ same gloss)

- **Text Normalization**: Language-specific handling (e.g., Uzbek apostrophe normalization)

The code is available at https://github.com/harapalb66/GLimmer.

5 Results and Analysis

5.1 Results

Table 1 shows aggregate results across all 15 languages. GLIMMER achieves 50.2% overall accuracy with strong ranking quality (nDCG: 0.804), placing fourth in the ADMIRE shared task competition (Pickard et al., 2025). Performance is higher for literal expressions (54.4%) than for idiomatic ones (46.3%), which is expected given that literal meanings are more directly grounded in visual evidence.

System	Acc ↑	ρ ↑	nDCG ↑
Shared Task Winner			
ITUNLP	**0.600**	—	**0.850**
GLIMMER (Our System)			
Overall	0.502	0.191	0.804
Idiomatic	0.463	0.187	0.778
Literal	0.544	0.194	0.835

Table 1: Aggregate results across 15 languages compared to the shared task winner.

Table 2 shows per-language performance. Portuguese-Brazil (66.7%), Russian (62.9%), and Slovenian (58.8%) achieve the highest accuracy, while Spanish-Ecuador (33.3%) and Igbo (37.4%) are most challenging.

Language	Acc	ρ	nDCG
Chinese	0.436	0.131	0.769
Georgian	0.496	0.135	0.791
Greek	0.543	0.310	0.831
Igbo	0.374	0.032	0.740
Kazakh	0.506	0.292	0.815
Norwegian	0.510	0.161	0.804
PT-BR	**0.667**	0.261	0.876
PT-PT	0.545	0.197	0.828
Russian	0.629	0.309	0.851
Serbian	0.479	0.161	0.784
Slovak	0.510	0.216	0.815
Slovenian	0.588	0.227	0.839
ES-EC	0.333	0.029	0.745
Turkish	0.484	0.118	0.800
Uzbek	0.425	0.289	0.774

Table 2: Per-language overall results. Best accuracy in bold.

5.2 Ablation Study

To assess the contribution of each component, we evaluate three variants on development data:

1. **Embed-only**: $score = score_{\text{embed}}$ (no LLM verification)
2. **Text-only**: $score_{\text{embed}} = sim_{\text{text}}$ (no CLIP)
3. **LLM-only**: Direct image-expression matching without gloss (no embeddings)

The following trends are observed across languages:

- CLIP vision path improves performance on visually distinctive cases
- LLM verification corrects embedding errors in subtle semantic distinctions
- Gloss-based grounding outperforms direct matching

Hyperparameter Tuning To determine the optimal fusion weights, we evaluated multiple configurations on development data. We found that assigning higher importance to textual captions ($\alpha = 0.6$) over visual features provided better semantic discrimination for abstract concepts. Similarly, prioritizing the LLM verification score ($\beta = 0.6$) over embedding similarity yielded the highest correlation across languages, as the reasoning capabilities of the LLM were crucial for correcting misalignments where embeddings conflated literal and idiomatic meanings.

5.3 Cross-Lingual Performance

We analyze performance across language families:

- **Romance** (PT-BR, PT-PT, ES-EC): Wide variance (33.3%-66.7%). Portuguese variants excel while Spanish-Ecuador struggles, possibly due to regional expression variations.
- **Slavic** (RU, SR, SK, SL): Strong overall (47.9%-62.9%), with Russian achieving the second-best accuracy. High Spearman correlations suggest good ranking quality.
- **Turkic** (KK, TR, UZ): Mixed results (42.5%-50.6%). Despite lower accuracy, Kazakh and Uzbek show surprisingly high correlations (0.289-0.292), indicating good relative ranking.

- **Other** (ZH, KA, EL, IG, NO): Greek performs exceptionally well (ρ=0.310, highest correlation), while Igbo is most challenging (37.4%), likely due to sparse representation in LLM pretraining corpora and culturally specific expressions with limited web image coverage.

The high nDCG scores (>0.74 across all languages), as computed via the official Codabench evaluation, demonstrate that GLIMMER produces reasonable rankings even when top-1 accuracy is modest, a valuable property for retrieval applications.

5.4 Error analysis

We examine a representative failure involving the Chinese compound 黑箱 ("black box"):

> 缺乏严格的程序性审查，导致很多加分通过黑箱操作等不正当的手段来获取。
>
> *"Lack of procedural review leads to bonus points obtained through black box operations and other **improper means**."*

Despite explicit corruption markers (不正当的手段, "improper means"), our system ranked images as shown in Table 3.

Table 3: Ranking failure for idiomatic

Image	System	Expected
Circuit cube	1st	–
Businessmen	5th	1st

Root cause: The LLM-generated gloss ("opaque operations") captured abstract semantics but missed pragmatic entailments: *human agency*, *institutional corruption*, *illicit gain*. This caused embeddings to prefer visually complex objects (circuits) over contextually appropriate scenes (businesspeople).

The *businessmen* image shows no lexical overlap with "opaque operations," yielding low text similarity and low CLIP similarity. Although LLM scoring (60% weight) recognized contextual fit, embedding scores (40%) had already created an insurmountable gap.

6 Limitations and Broader Impact

6.1 Limitations

Our approach depends on LLM-based gloss generation and verification. While gloss caching improves efficiency, future work will explore distilling gloss generation into smaller or open models. Additionally, gloss quality may vary for culturally specific idioms, and errors at this stage can propagate through the pipeline. Incorporating explicit uncertainty quantification mechanisms could improve system transparency and reliability (Ni et al., 2025). The dual-modality architecture requires both captions and images, limiting applicability to caption-free scenarios.

6.2 Broader Impact

Improved idiom grounding benefits multilingual retrieval, education, and cross-cultural communication tools. However, biases present in web imagery or LLM training data may influence rankings, particularly for under-resourced languages. Our system reflects patterns learned from existing data, which may not capture the full diversity of idiomatic usage across cultures.

7 Conclusion

We presented GLIMMER, a hybrid system for ranking images by multiword expression fit across 15 languages. Our gloss-based architecture provides a stable semantic anchor enabling cross-lingual transfer, while hybrid retrieval-reasoning scoring balances efficiency and semantic precision. Key findings indicate that i) gloss-based representations enable multilingual transfer without parallel data, ii) hybrid retrieval-reasoning architectures outperform embedding-only approaches, and iii) the integration of textual and visual modalities improves robustness to caption quality.

GLIMMER achieves 50.2% overall accuracy with strong ranking quality (nDCG: 0.804) across 15 typologically diverse languages, demonstrating that explicit semantic anchoring is an effective strategy for multimodal idiom understanding.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, and 1 others. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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*.

Paul-Ambroise Duquenne. 2024. *Sentence Embeddings for Massively Multilingual Speech and Text Processing*. Ph.D. thesis, Sorbonne Université.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023b. Dual-alignment pre-training for cross-lingual sentence embedding. *arXiv preprint arXiv:2305.09148*.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, and 1 others. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

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. *arXiv preprint arXiv:2503.15358*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23.

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 e-prints*, pages arXiv–2204.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational linguistics: NAACL 2022*, pages 2715–2726.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 174–179.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

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*.