Picture This: Disambiguating Literal and Figurative Idioms with Text and Images

Adi Carmi 305690216, Omer Hanani 315003046

August 24, 2025

1 Introduction

Disambiguating figurative from literal meanings of Potentially Idiomatic Expressions (PIEs) such as ”blind as a bat”, “kick the bucket” or ”raise the roof”, is a persistent challenge in natural language understanding (NLU). Idioms are often non-compositional, their interpretation depends on context and cannot be inferred from individual words. For example, *“He kicked the bucket in the yard”* is literal, while *“He finally kicked the bucket”* figuratively represents death. Misclassification of figurative vs. literal usage undermines tasks such as translation, image generation, and dialogue systems. Even large language models struggle on this task. Visual context offers a promising solution: images can supply disambiguating cues, echoing the adage that *“a picture is worth a thousand words”*. We propose a multimodal approach that grounds idioms in literal and figurative image anchors, leveraging visual cues to aid disambiguation. Specifically, we fine-tune MobileCLIP [5], a lightweight CLIP variant, with paired idiom–image references. To strengthen general representations and reduce overfitting, we added an auxiliary image– text retrieval objective, inspired by COCO-style contrastive training. Our analysis demonstrates the potential of visual grounding and joint training objectives for improving PIE disambiguation.

1.1 Related Work

The MAGPIE corpus [4] introduced the first large, sense-annotated resource for idiomatic expressions, enabling systematic study of idiomaticity. Drawn from the British National Corpus, it has become the standard benchmark for idiom disambiguation. Early work leveraged MAGPIE to develop baselines: Zeng and Bhat [8] used semantic compatibility for idiom detection, while Dankers et al. [3] showed that Transformer-based NMT systems still fail on non-compositional idioms. Adewumi et al. [2] later released PIE-English with ten idiom classes, but its inflated scores (F1 $\approx$ 0.95 with BERT) largely reflect idiom class imbalance, limiting comparability to MAGPIE. More recently, Phelps et al. [6] reported that RoBERTa-large achieves 91.9% accuracy on MAGPIE, whereas state-of-the-art LLMs such as GPT-4-turbo reach only $\approx$ 81% zero-shot, with open source models performing worse. These findings highlight MAGPIE’s value as benchmark for idiomaticity.

In parallel, multimodal representation learning has advanced rapidly. CLIP [7] demonstrated that aligning image and text embeddings at scale yields powerful zero-shot transfer and robust generalization. MobileCLIP [5] extends these principles to resource-constrained settings through lighter architectures and multimodal reinforcement, enabling efficient deployment on mobile devices. Beyond encoders, recent advances in image generation (e.g., the FLUX pipeline in Hugging Face Diffusers [1]) and organizational tools like Hugging Face Sheets facilitate scalable creation of paired literal and figurative idiom imagery, providing the visual anchors needed for multimodal idiom disambiguation.

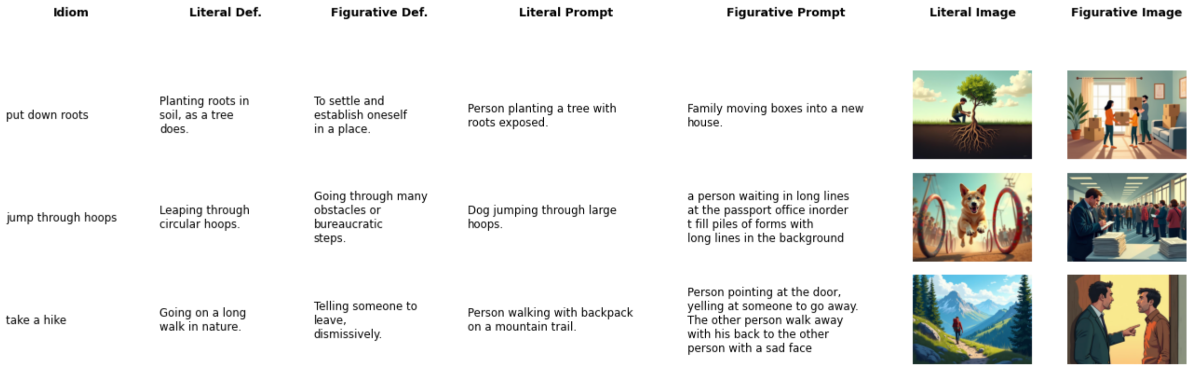
2 Methodology

2.1 General Approach

The core approach of the project was to align the idiom span embeddings with literal and figurative image embeddings in a shared multimodal space. Idiomatic expressions are captured as token-level span embeddings extracted from sentences, while reference images are encoded into fixed feature vectors. During fine-tuning, the text encoder is adjusted so that span embeddings move closer to their correct visual anchors. In parallel, auxiliary objectives shape the embedding space to separate literal span embeddings from figurative usages in order to preserve generalization. This design allows the model to learn idiom-specific disambiguation while maintaining broader multimodal alignment.

2.2 Datasets

Custom Idiom Images Dataset: We constructed this dataset to provide explicit visual grounding for idiomatic meaning serving as reference anchors for disambiguation. It contains 36 idioms sampled from the MAGPIE corpus. GPT-4.1 was used to generate definitions and representative image prompts for both literal and figurative meanings of each idiom. Hugging Face Sheets was then used to produce figurative and literal anchor images with the FLUX DEV model [1]. For each idiom, two reference images were generated, one literal and one figurative.



Idiom	Literal Def.	Figurative Def.	Literal Prompt	Figurative Prompt	Literal Image	Figurative Image
put down roots	Planting roots in soil, as a tree does.	To settle and establish oneself in a place.	Person planting a tree with roots exposed.	Family moving boxes into a new house.		
jump through hoops	Leaping through circular hoops.	Going through many obstacles or bureaucratic steps.	Dog jumping through large hoops.	a person waiting in long lines at the passport office inorder t fill piles of forms with long lines in the background		
take a hike	Going on a long walk in nature.	Telling someone to leave, dismissively.	Person walking with backpack on a mountain trail.	Person pointing at the door, yelling at someone to go away. The other person walk away with his back to the other person with a sad face		

Figure 1: A sample of the Custom Idiom Image dataset.

MAGPIE Corpus [4]: A large, sense-annotated dataset of potentially idiomatic expressions (PIEs) with span-level labels indicating literal or figurative usage. For each idiom in the Custom Idiom Images Dataset, we curated the corresponding MAGPIE instances, yielding 1,180 labeled examples.

MS COCO Captions (Lin et al.): A subset of 1,000 paired COCO images and captions was used as auxiliary supervision for image-text retrieval training.

2.3 Architecture

MobileCLIP (S0 variant) was chosen as the base model. It is designed with lightweight architectures to reduce computational cost while maintaining competitive performance on multimodal tasks. During training, the visual encoder is frozen while only the text encoder is fine-tuned.

2.3.1 Span Extraction and Image Embeddings

Span Extraction. In MAGPIE, idioms are annotated at the *word level*, specifying start and end indices within each sentence. Since MobileCLIP uses a subword tokenizer, words may split into multiple tokens (e.g., "kicked" → `kick`, `##ed`).To align the MAGPIE annotations with the model input, a binary `span_token_mask` is constructed by: (1) tokenizing the entire sentence, (2) Identify all tokens corresponding to the words annotated as idiomatic, (3) Mark these token positions with 1, and all others with 0 , and (4) excluding special tokens (e.g., `[CLS]`, `[SEP]`). For example, *"He kicked the bucket"* tokenizes as `[CLS], He, kick, ##ed, the, bucket, [SEP]`. If the idiom span is *"kicked the bucket"*, the mask is `[0, 1, 1, 1, 1, 0]`. During encoding, embeddings for tokens with mask 1 are extracted and mean-pooled to form a single *idiom span embedding*, representing the idiom in its sentence context.

Image Embeddings. Each idiom is paired with two reference images, illustrating its *literal* and *figurative* senses. Images are resized, normalized, and passed through the MobileCLIP image encoder to obtain fixed-dimensional feature vectors. These embeddings are computed once at setup and remain frozen, serving as semantic anchors during training. The text encoder (span embeddings) is fine-tuned to align with these fixed semantic anchors via contrastive learning.

2.3.2 Loss components:

The following losses were leverage to jointly train the model on the Idiom Disambiguation and COCO retrieval tasks.

1. **Span-to-image cross-entropy**: *Algorithm*: InfoNCE (Information Noise-Contrastive Estimation), a contrastive learning objective implemented via cross-entropy. *Mechanism*: cosine similarity between span s_i and images v_j is normalized with temperature τ. The loss is
$$L_{\text{span-img}} = -\frac{1}{N}\sum_i \log \frac{\exp(\text{sim}(s_i, v_i)/\tau)}{\sum_j \exp(\text{sim}(s_i, v_j)/\tau)}.$$
Contribution: aligns each idiom span with its paired image and uses all other images as negatives, , following the same principle as CLIP training.

2. **Supervised contrastive loss (Hierarchical)**: *Algorithm*: Weighted supervised contrastive (SupCon) objective. *Mechanism*: For each anchor span i, positives are spans sharing the same usage label; these positives are weighted to distinguish *same idiom + same usage* from *different idiom + same usage.* The per-anchor loss is
$$\ell_i = -\frac{\sum_j w_{ij}\,\mathbf{1}_{\text{pos}(i,j)}\,\log p(j\,|\,i)}{\sum_j w_{ij}\,\mathbf{1}_{\text{pos}(i,j)} + \varepsilon}, \qquad p(j\,|\,i) = \frac{\exp\big(\text{sim}(s_i, s_j)/\tau\big)}{\sum_{a\neq i}\exp\big(\text{sim}(s_i, s_a)/\tau\big)},$$
with cosine similarity $\text{sim}(s_i, s_j) = \frac{\langle s_i, s_j\rangle}{\|s_i\|\,\|s_j\|}$ and weights
$$w_{ij} = \begin{cases} w_{\text{same-idiom}} & \text{if same idiom + same usage,} \\ w_{\text{cross-idiom}} & \text{if different idiom + same usage,} \\ 0 & \text{otherwise,} \end{cases} \qquad w_{\text{same-idiom}} > w_{\text{cross-idiom}} \geq 0.$$
We set $w_{\text{same-idiom}} = 1, w_{\text{cross-idiom}} = 0.5$. The batch loss is $L_{\text{supcon}} = \frac{1}{N}\sum_i \ell_i$, *Contribution*: Produces tight intra-idiom clusters for the same label (strong pull) while still encouraging broader grouping across idioms for the same label (weak pull), improving discrimination without collapsing all literal or all figurative spans into a single global cluster.

3. **KL-divergence loss**: *Algorithm*: Kullback–Leibler divergence. *Mechanism*: distills sentence embedding predictions from a frozen MobileCLIP teacher. With distributions p (teacher) and q (student):
$$L_{\mathrm{KL}} = \sum_j p_j \log \frac{p_j}{q_j}.$$
Contribution: stabilizes fine-tuning and prevents catastrophic forgetting by preventing the model from diverging to far from its pretrained setting.

4. **Pairwise margin loss**: *Algorithm*: hinge ranking loss. *Mechanism*: enforces margin m between positive and negative similarities:
$$L_{\mathrm{pair}} = \max(0, m - (s_{pos} - s_{neg})).$$
Contribution: directly contrasts the correct vs. opposite image, ensuring a minimal similarity gap.

5. **Auxiliary COCO retrieval loss**: *Algorithm*: symmetric InfoNCE. *Mechanism*:Contrastive training applied stochastically to COCO images and captions with bidirectional objectives:
$$L_{\mathrm{COCO}} = \tfrac{1}{2}(L_{\mathrm{img}\to\mathrm{txt}} + L_{\mathrm{txt}\to\mathrm{img}}).$$
Contribution: preserves the sentence level semantics, reducing the impact of overfitting to idiom-specific images.

Total Training Objective: The final loss is a weighted combination of all components:

$$L_{\mathrm{total}} = L_{\text{span-img}} + \lambda_{\mathrm{supcon}} L_{\mathrm{supcon}} + \lambda_{\mathrm{KL}} L_{\mathrm{KL}} + \lambda_{\mathrm{pair}} L_{\mathrm{pair}} + \lambda_{\mathrm{retrieval}} L_{\mathrm{COCO}},$$

where $\lambda_{\mathrm{supcon}}$, λ_{KL}, λ_{pair}, and $\lambda_{\mathrm{retrieval}}$ are hyperparameters controlling the contribution of each auxiliary loss.

Ablation Study To better understand the contribution of each loss term, we conducted a series of ablation experiments. In each run, one or more components (e.g., supervised contrastive, KL-distillation, COCO retrieval, or pairwise margin) were removed, keeping the others intact.

2.4 Platform and Training Process

Experiments were conducted in Google Colab using Python, with Pandas for preprocessing, PyTorch Lightning for training, HuggingFace for tokenization and for loading MobileCLIP checkpoints, scikit-learn for evaluation, and matplotlib for visualization. MAGPIE and MS COCO were integrated via custom PyTorch DataLoaders. Models were trained with AdamW, a cosine annealing learning rate schedule, mixed-precision (16-bit), and a batch size of 16, for 10 epochs on a single GPU. Runs were executed on free NVIDIA T4 GPUs (16GB). Each model experiment required approximately 7 minutes to complete training. GPU memory limitations constrained the batch size, and caused occasional failures.

3 Experimental Results

3.1 Settings

Data Splits: MAGPIE idiom instances were divided into training (944) and validation (236) sets. For COCO retrieval, 500 images were used for training and 500 for validation.

Hyperparameters: Unless otherwise noted, loss weights were set to $(\lambda_{\text{supcon}}, \lambda_{\text{KL}}, \lambda_{\text{pair}}, \lambda_{\text{retrieval}}) = (0.5, 0.1, 2.0, 1.0)$, with pairwise margin $m = 0.2$ and contrastive temperature $\tau = 0.07$. The COCO retrieval loss was activated stochastically with probability $p = 0.1$ per step.

Evaluation Protocol: For each run, we logged the losses every step and the total loss at every step and epoch. The best model was selected based on minimum training total loss. Models were evaluated on two tasks. For idiom usage classification, we used the cosine similarity between the idiom span embedding and its two reference images (literal and figurative). The predicted label was assigned according to which image had the higher similarity score, and performance was reported using accuracy, precision, recall, and F1 (both per-class and weighted). For COCO image–text retrieval, we followed the standard CLIP-style evaluation: given a batch of images and captions, we computed pairwise cosine similarities between embeddings, ranked them, and reported Recall@1 and Recall@5 for both directions for both image-to-text (I2T) and text-to-image (T2I).

Reproducibility: Random seeds were set to 42 across Python, NumPy, and PyTorch, with deterministic cuDNN enabled. Library versions: PyTorch 2.8.0, PyTorch Lightning 2.5.3, HuggingFace Transformers 4.55.2

Ablation Configurations. To study the role of auxiliary objectives, we trained seven ablation variants. Table 1 summarizes which loss terms were active in each setting. The baseline corresponds to the original pretrained MobileCLIP model. All other variants included the span-to-image contrastive loss as the core idiom disambiguation objective, with objectives selectively disabled: Pair Margin (pairwise margin only), Pair Margin Span2Span (pairwise margin + Supervised contrastive loss), No KL (drops KL-divergence), No COCO (drops COCO retrieval), No Span2Span (drops Supervised contrastive loss), and Full (all objectives).

Configuration	Span-to-Image	Span-to-Span	KL-Div	Pair Margin	COCO Retrieval
Baseline (pretrained CLIP)	x	x	x	x	x
Pair_Margin	v	x	x	v	x
Pair_Margin_Span2Span	v	v	x	v	x
No_KL	v	v	x	v	v
No_COCO	v	v	v	x	v
No_Span2Span	v	x	v	v	v
Full	v	v	v	v	v

Table 1: Ablation study configurations. All fine-tuned variants include span-to-image alignment; auxiliary losses are selectively removed to isolate their contributions.

3.2 Results

Fine-tuning MobileCLIP with idiom-grounded objectives raised idiom classification accuracy from 0.51 (baseline) to nearly 0.90 across the fine-tuned settings (Table 2), while only moderately affecting retrieval alignment. The Span-to-image alignment with the pairwise margin loss was the main driver, with figurative F1 0.92, and Literal F1 0.86. Adding the span-to-span contrastive learning resulted in the same F1 values, but causes the retrieval to drop. KL-regularization had little effect. For COCO retrieval, (Table 3), the pretrained baseline achieved the best Recall@5 (≈0.77–0.78), whereas fine-tuned models were slightly lower (≈0.61–0.76). The auxiliary COCO loss helped preserve and in some cases even improve alignment over the

Ablation	Accuracy	F1-Literal	F1-Figurative	Weighted F1
Baseline	0.508	0.408	0.580	0.518
Pair margin	**0.903**	**0.862**	**0.925**	**0.902**
Pair margin + span2span	0.903	0.862	0.925	0.902
No KL	0.898	0.857	0.921	0.898
No COCO	0.894	0.850	0.918	0.894
No span2span	0.886	0.840	0.911	0.885
Full model	0.886	0.842	0.910	0.886

Table 2: Ablation study: Idiom classification results.

Ablation	COCO I2T@1	COCO I2T@5	COCO T2I@1	COCO T2I@5
Baseline	0.266	**0.780**	0.234	**0.774**
Pair margin	0.252	0.744	0.232	0.654
Pair margin + span2span	0.242	0.734	0.176	0.618
No KL	0.258	0.766	0.228	0.746
No COCO	0.250	0.752	0.176	0.646
No span2span	**0.272**	0.756	0.236	0.730
Full model	0.246	0.770	**0.256**	0.766

Table 3: Ablation study: COCO retrieval results.

baseline. Removal of the auxiliary COCO loss caused the retrieval to drop, confirming its role in preventing multimodal forgetting.

3.3 Error Analysis

To better understand model behavior, we analyzed the fine-tuned prediction errors against gold labels. We define confidence as the similarity margin between a span embedding and its literal vs. figurative image anchors:

$$\text{margin}(i) = \left|\text{sim}(s_i, v_{\text{lit}}) - \text{sim}(s_i, v_{\text{fig}})\right|,$$

where sim is cosine similarity. Larger margins indicate stronger model preference, making this a useful proxy for ranking errors. The most common high-confidence errors were literal usages misclassified as figurative (e.g., *add fuel to the fire* used literally), reflecting both dataset imbalance (more figurative than literal samples) and the ambiguity of idioms with plausible literal readings (e.g., *turn the corner*, *in hot water*, *rock the boat*), that are often carry strong figurative interpretations but remain common in literal use.

Representative Cases. Tables 4 and 5 show illustrative examples. Table 4 highlights errors corrected by fine-tuning (e.g., *break the bank*, *apples and oranges*), while Table 5 shows remaining high-confidence mistakes, dominated by literal usages predicted as figurative.

At the per idiom level, our method improved performance for 14 of 32 idioms that were examined (44%), with the largest gains in cases where the baseline completely failed (e.g., "break the bank" 0.000 → 0.462, "in broad daylight" 0.000 → 0.500, "watering hole" and "in a rut" both reaching 1.000 F1). Idioms with both literal and figurative senses also benefited, such as "cut corners" (0.091 → 0.474) and "turn the corner" (0.422 → 0.557), suggesting that supervision helps resolve ambiguity. However, performance remained unchanged for 13 idioms (41%) and declined modestly for 5 (16%), including "jump through hoops" (0.667 → 0.400) and "have a ball" (0.788 → 0.462).

Idiom	Sentence	Gold	Conf.
Break the Bank	At under £14,500 it won't **break the bank** either.	F	0.349
All Wet	Had to spend the War in England, of course—ghastly place, **all wet** and miserable.	L	0.303
Rock the Boat	No one will want to **rock a boat** that has almost capsized.	F	0.301
Apples and Oranges	Allow 100 calories from the half pint of skimmed milk, and another 100 for your daily **apple and orange**.	L	0.300
Rock the Boat	Nobody was willing to **rock the boat** with a rescue operation.	F	0.297

Table 4: Examples of baseline errors corrected after training (F = figurative, L = literal).

Idiom	Text	Pred.	Gold	Conf.
Add fuel to the fire	She sat up to **add more fuel to the fire**.	F	L	0.361
Fly off the handle	The head **flew off the handle** and cracked our only mirror.	F	L	0.329
Make waves	Paul floated a toy boat and beat the water to **make waves**.	F	L	0.258
Put your foot down	You really have to **put your foot down** sometimes.	F	L	0.250
Put down roots	Mosses nor lichens **put down roots**, so organic material remains superficial.	F	L	0.236

Table 5: Remaining high-confidence misclassifications (Pred. F = figurative, L = literal).

Where Table 4. shows significant improvement over the MobileClip baseline, table 5. demonstrates that ambiguous idioms remain challenging: literal senses are systematically harder, often over-predicted as figurative, and dataset imbalance further inflates accuracy for idioms restricted to a single sense without guaranteeing generalization.

4 Discussion and Conclusion

This work introduced a novel application of multimodal grounding for idiom understanding by fine-tuning MobileCLIP with idiom-specific literal and figurative images. To our knowledge, no prior study has attempted to anchor idiomatic expressions in visual space, making this approach both original and promising. Fine-tuning substantially improved performance, raising idiom accuracy from about 0.51 in the baseline to nearly 0.90 across the fine-tuned setting, with weighted F1 exceeding 0.90. The largest gains arose from combining span-to-image alignment with pairwise margin and supervised contrastive losses. Auxiliary COCO retrieval further acted as a regularizer, preserving multimodal alignment, preventing overfitting to the relatively small idiom dataset, even improving COCO T2I@1 retrieval perfomance over the baseline.

Despite these gains, several challenges remain. Literal prediction was consistently weaker, as many errors involved literal spans being pulled toward figurative interpretations. Ambiguous idioms also proved difficult, and dataset imbalance inflated metrics for idioms with limited or single-sense examples. These limitations underscore the significance of the positive results: strong improvements were achieved despite data scarcity and limited resources. We conjecture that weaker outcomes partly reflect noise in image-anchor generations, which may fail to capture idiomatic nuance, and partly the small test sizes for several idioms (as few as 3–5 examples). Importantly, our per idiom analysis highlights that overall F1 can be misleading on MAGPIE, since global averages obscure wide variation across idioms. Future progress therefore requires not only improved multimodal architectures but also balanced datasets and evaluations that foreground per-idiom distributions, ensuring gains are driven proportionately across all idioms.

Overall, these findings demonstrate that visual grounding can substantially enhance idiom disambiguation, correcting systematic baseline errors and paving the way for richer multimodal approaches to figurative language. In particular, visual anchors help resolve ambiguity, while supervised contrastive learning strengthens embedding structure. Error analyses further revealed systematic patterns tied to idiom frequency and distribution, reinforcing the need for methodological advances in both modeling and dataset design.

This project provides an early but significant proof of concept: linking idiomatic text with visual anchors improves robustness where textual cues alone fall short. Future directions include scaling to larger and more balanced idiom corpora, exploring dynamic loss weighting for multitask optimization (e.g., via Optuna), and integrating richer visual features. Priorities involve improving literal recognition through data augmentation and targeted error analysis, while alternative losses such as triplet, angular margin, or center loss may further strengthen clustering and mitigate imbalance. Robustness testing, interpretability methods (e.g., attention maps, embedding visualizations), and integration with word sense disambiguation datasets (e.g., PolCLIP, ACL 2024) also present promising avenues. More broadly, this framework may extend to metaphors, sarcasm, and other figurative phenomena, opening a new direction for multimodal NLP research.

5 Code

Git Link: https://github.com/adicarmi/nlp_idioms

References

[1] Flux text-to-image generation via hugging face diffusers (`fluxpipeline`). Hugging Face documentation, 2025. `https://huggingface.co/docs/diffusers/en/api/pipelines/flux`.

[2] Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. Potential idiomatic expression (pie)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*, 2022.

[3] Verna Dankers, Jack Lucas, and Ivan Titov. Can transformers be too compositional? analysing idioms in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1151–1165, Dublin, Ireland, 2022.

[4] Hessel Haagsma, Johannes Bjerva, and Malvina Nissim. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of LREC*, 2020.

[5] Fartash Faghri Raviteja Vemulapalli Oncel Tuzel Pavan Kumar Anasosalu Vasu, Hadi Pouransari. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[6] Patrick Phelps, Michael Kim, Glorianna Jagfeld, and Ngoc Thang Vu. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the 20th Workshop on Multiword Expressions (MWE)*, pages 181–192, St. Julians, Malta, 2024. Association for Computational Linguistics.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, 2021.

[8] Changsheng Zeng and Suma Bhat. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1359–1374, 2021.