

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



Adi Carmi 305690216, Omer Hanani 315003046
August 24, 2025

Introduction – The Challenge

- Disambiguating figurative from literal meanings of Potentially Idiomatic Expressions (PIEs), is a persistent challenge in natural language understanding (NLU)



Introduction – The Challenge

- Idioms are often non-compositional, their interpretation depends on context and cannot be inferred from individual words.
- Misclassification of figurative vs. literal usage undermines tasks such as translation and image generation. Even LLM struggle on this task.

Generate a figurative interpretation image of the potential idiom phrase flying off the handle.

Image created



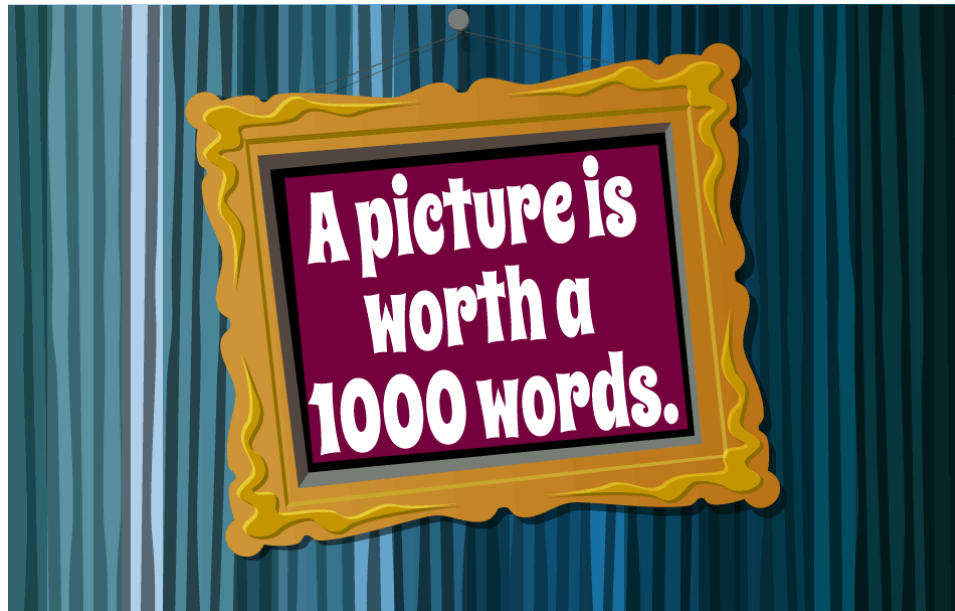
Generate a literal interpretation image of the potential idiom phrase flying off the handle.

Image created



Introduction – Our Idea

- Using Visual context for grounding Images can supply disambiguating cues



Project Objectives

- **Investigate multimodal grounding for improving PIE disambiguation.**

Test whether linking idiomatic expressions to images can improve disambiguation between *literal* and *figurative* usage.

Methodology Overview

- We propose a multimodal approach that grounds idioms in literal and figurative image anchors, leveraging visual cues to aid disambiguation.
- We fine-tuned MobileCLIP (S0 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.

Datasets

- Custom Idiom-Image dataset: 36 idioms sampled from MAGPIE, with literal + figurative images, generated with FLUX DEV model



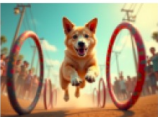



Idiom	Literal Def.	Figurative Def.	Literal Prompt	Figurative Prompt	Literal Image	Figurative Image
put down roots	Planting roots in soil, as a tree does.	To settle and establish oneself in a place.	Person planting a tree with roots exposed.	Family moving boxes into a new house.		
jump through hoops	Leaping 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 in order to 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.

Datasets

- For each idiom in the Custom Idiom Images Dataset, we curated the corresponding MAGPIE instances(1,180 labeled examples)

	sentence	annotation	idiom	usage
0	The boy pulled himself together so hastily tha...	0 1 1 ...	a hair's breadth	figurative
1	Martin Rosenbaum, the family adviser, added : ...	0 ...	a hair's breadth	figurative
2	The cutting edge of his sword was a hair's - b...	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0	a hair's breadth	figurative

- MS COCO: 1k image-caption pairs, used as auxiliary supervision for image-text retrieval training

Idiom Span Extraction & Embedding

In MAGPIE Idioms are marked at the word level (start/end indices), but MobileCLIP uses subword tokenization (e.g., kicked → kick, ##ed).

- **Solution – span mask:**

- Tokenize the full sentence with MobileCLIP.
- Mark idiom tokens with **1**, others with **0**.
- Exclude special tokens (e.g., [CLS], [SEP]).

- **Example:**

“He kicked the bucket” → tokenizes as [CLS], He, kick, ##ed, the, bucket, [SEP]. If the idiom span is “kicked the bucket”, the mask:
[0, 1, 1, 1, 1, 0].

- **Embedding:** Masked token embeddings are mean-pooled → single **idiom span embedding** (captures idiom in context).

Image Embeddings

- Each idiom paired with two reference images:
 - Literal sense
 - Figurative sense
- Images preprocessed: resized + normalized.
- Encoded by MobileCLIP image encoder → fixed-dimensional vectors.
- Embeddings are precomputed and frozen (not updated).

Multi-loss objective

Losses were leverage to jointly train the model on the Idiom Disambiguation and COCO retrieval tasks:

- Span-to-Image: InfoNCE loss
- Hierarchical Supervised Contrastive loss
- KL-divergence distillation loss
- Pairwise margin loss
- Auxiliary COCO retrieval loss

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

Where $\zeta \in \{0, 1\}$ indicates the stochastic inclusion of the COCO retrieval loss per batch

Span-to-Image Loss

- **Algorithm:** InfoNCE (contrastive cross-entropy).
- **Mechanism:** Idiom span embedding aligned with its paired image; all other images in batch act as negatives.
- **Contribution:** Provides direct supervision for literal vs. figurative grounding.
- This is standard in CLIP-style models, but here spans are the idiom spans (not whole sentences).

$$L_{\text{span-img}} = -\frac{1}{N} \sum_i \log \frac{\exp(\text{sim}(s_i, v_i)/\tau)}{\sum_j \exp(\text{sim}(s_i, v_j)/\tau)}$$

Where s_i is the idiom span embedding, and v_j is the image embedding, normalized with temperature τ

Hierarchical Supervised Contrastive Loss

- **Algorithm:** Weighted supervised contrastive (SupCon/span2span)
- **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. Strong pull for *same idiom + same usage*, weaker pull for *different idiom + same usage*.
- **Contribution:** Tightens intra-idiom clusters while still grouping across idioms.

$$p(j \mid i) = \frac{\exp(\text{sim}(s_i, s_j)/\tau)}{\sum_{a \neq i} \exp(\text{sim}(s_i, s_a)/\tau)}, \quad \text{sim}(s_i, s_j) = \frac{s_i^\top s_j}{\|s_i\| \|s_j\|}.$$

$$\alpha_{ij} = \begin{cases} w_{\text{same}} & \text{if (same idiom) \& (same usage)} \\ w_{\text{cross}} & \text{if (different idiom) \& (same usage)} \\ 0 & \text{otherwise} \end{cases}$$

We set $w_{\text{same}} = 1$ $w_{\text{cross}} = 0.5$

Per-anchor loss:

$$\ell_i = - \frac{\sum_{j \neq i} \alpha_{ij} \log p(j \mid i)}{\sum_{j \neq i} \alpha_{ij} + \varepsilon}$$

Total batch loss:

$$L_{\text{supcon}}^{\text{hier}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \ell_i, \quad \mathcal{I} = \{i : \sum_{j \neq i} \alpha_{ij} > 0\}.$$

KL-divergence distillation Loss

- **Algorithm:** KL divergence (knowledge distillation).
- **Mechanism:** Matches student (fine-tuned model) output distribution to frozen teacher (pretrained MobileCLIP).
- **Contribution:** Stabilizes training, avoids catastrophic forgetting by preventing the model from diverging too far from its pretrained setting

$$L_{\text{KL}} = \sum_j p_j \log \frac{p_j}{q_j}$$

Pairwise Margin Loss

- **Algorithm:** Hinge ranking loss.
- **Mechanism:** Enforces margin m between positive and negative similarities.
- **Contribution:** directly contrasts the correct vs. opposite image, ensuring a minimal similarity gap

$$L_{\text{pair}} = \max(0, m - (s_{\text{pos}} - s_{\text{neg}}))$$

Where s_{pos} is the idiom span embedding and correct image embedding similarity, and s_{neg} is the idiom span embedding and opposite image embedding similarity

Auxiliary COCO Retrieval Loss

- **Algorithm:** Symmetric InfoNCE.
- **Mechanism:** Bidirectional contrastive training on COCO images and captions.
- **Contribution:** Provides broad multimodal alignment, reduces overfitting to idiom-specific images

$$L_{\text{COCO}} = \frac{1}{2} \left(L_{\text{img} \rightarrow \text{txt}} + L_{\text{txt} \rightarrow \text{img}} \right)$$

Training Setup

- Images serves as semantic anchors during training.
- The text encoder (idiom spans) is fine-tuned to align with these anchors via contrastive learning and auxiliary losses.
- Platform: Colab, PyTorch Lightning
- Single T4 GPU (16GB)
- 10 epochs, batch size 16
- AdamW optimizer + cosine annealing
- Mixed-precision training (16-bit)
- Checkpointing on loss
- About 7 minutes to complete training for each model configuration

Training Setup

- MAGPIE idiom instances: 80/20 split, training (944) and validation (236)
- For COCO retrieval, 500 images were used for training and 500 for validation.
- Hyperparameters: Unless otherwise noted, loss weights were set to:
 - $(\lambda_{\text{supcon}}, \lambda_{\text{KL}}, \lambda_{\text{pair}}, \lambda_{\text{retrieval}}) = (0.5, 0.1, 2.0, 1.0)$
 - pairwise margin $m = 0.2$
 - contrastive temperature $\tau = 0.07$.
 - COCO retrieval loss was activated stochastically with probability $p = 0.1$ per step

Evaluation Protocol

Training

- Logged loss per step and total loss per epoch
- Selected best model based on *minimum training total loss*

Task 1 – Idiom Usage Classification

- Compared span embedding with literal vs. figurative image embeddings
- Predicted label = image with higher cosine similarity
- Metrics: Accuracy, Precision, Recall, F1 (per-class + weighted)

Task 2 – COCO Image–Text Retrieval

- Standard CLIP-style evaluation
- Computed pairwise cosine similarities between images & captions
- Ranked similarities → reported Recall@1, Recall@5
- Both directions: Image→Text (I2T) and Text→Image (T2I)

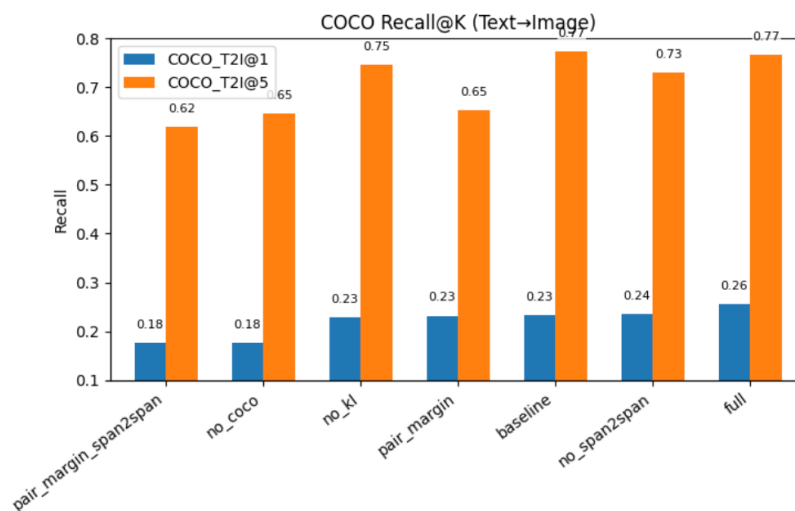
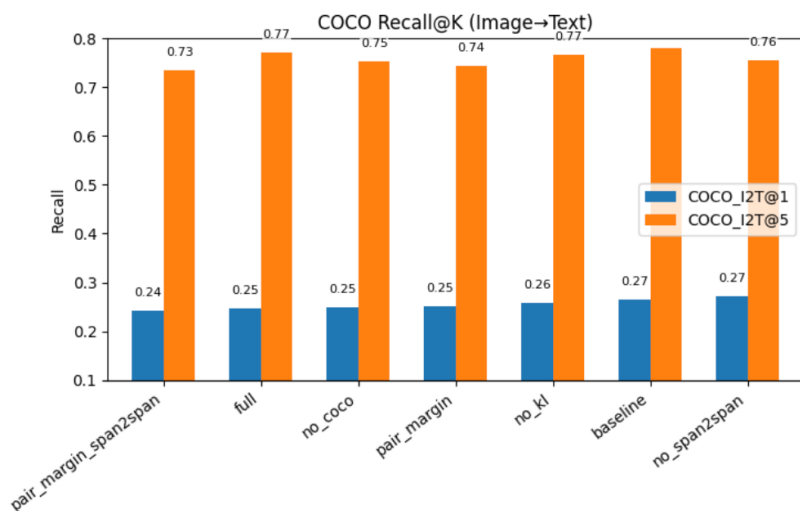
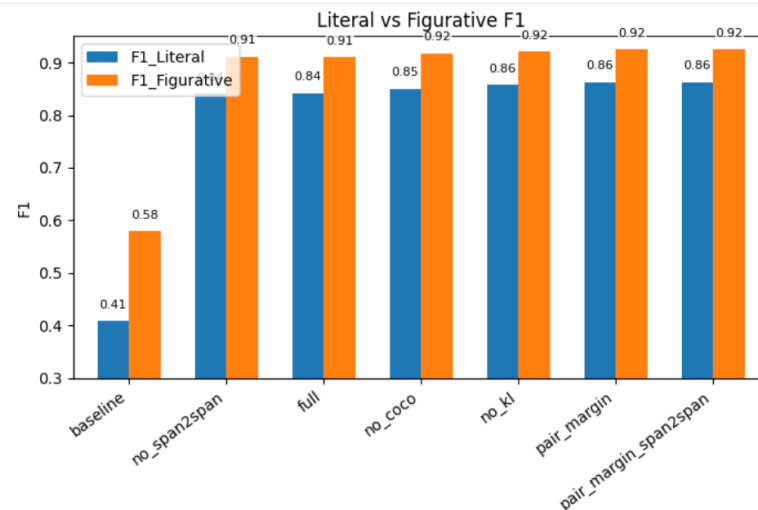
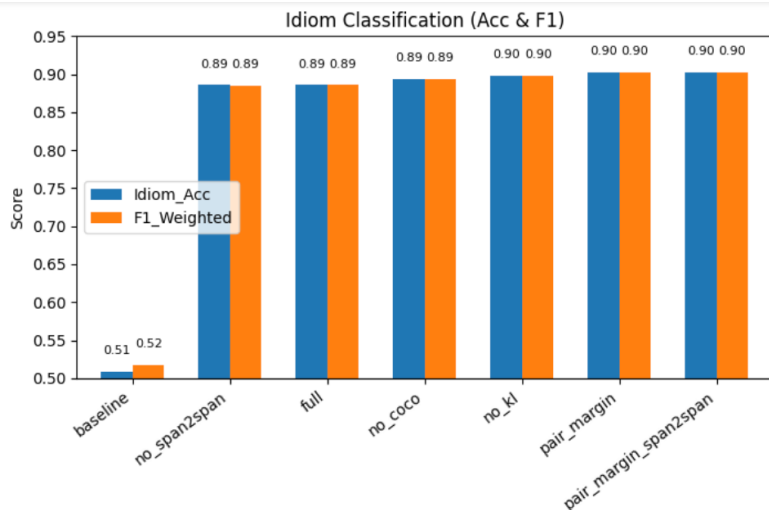
Ablation Study

- 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(Span2Span), KL-distillation, COCO retrieval, or pairwise margin) were removed, keeping the others intact.
- Baseline: pretrained MobileCLIP
- Variants: removed individual loss terms
- Evaluated idiom classification & COCO retrieval
- Main driver: span-to-image loss in all variants, as the core idiom disambiguation objective

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.

Results



Results

Idiom Classification

- Accuracy improved: **0.51** → **~0.90**
- Figurative F1: **0.92**
- Literal F1: **0.86**

Key Loss Contributions

- Span-to-image + Pairwise margin → main driver of gains
- Hierarchical SupCon (span2span) → same results as Pairwise margin but dropped the Retrieval Recall.
- KL-regularization → minor effect
- COCO loss → stabilized multimodal alignment

COCO Retrieval

- Baseline Recall@5: **0.77–0.78**
- Fine-tuned models: **0.71–0.76**
- Auxiliary COCO loss = prevents forgetting, preserves the sentence level semantic

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.

Error Analysis

- Fine-tuning corrected many baseline mistakes
- Our method improved performance for 14 of 32 idioms that were examined (44%)

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).

Error Analysis

- Most common errors: literal misclassified as figurative, reflecting:
 - Bias: dataset imbalance (more figurative examples)
 - 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

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).

Error Analysis

- Our method improved performance for 14 of 32 idioms that were examined (44%):
 - Strongest gains on idioms where the baseline failed (e.g., *break the bank*, *in broad daylight*).
 - Literal/figurative idioms improved (e.g., *cut corners*, *turn the corner*).
 - 41% unchanged, 16% modest decline (e.g., *jump through hoops*, *have a ball*).

Idiom	Total	Lit	Fig	F1_Baseline	F1_Trained
walk the plank	1	1	0	0.000	0.000
fly off the handle	3	1	2	0.400	0.400
jump through hoops	3	1	2	0.667	0.400
take the cake	3	2	1	0.250	0.400
put down roots	4	1	3	0.333	0.429
in hot water	9	7	2	0.181	0.438
add fuel to the fire	5	1	4	0.444	0.444
break the ice	5	1	4	0.583	0.444
make waves	6	1	5	0.455	0.455
put your foot down	6	1	5	0.455	0.455
break the bank	7	0	7	0.000	0.462
have a ball	7	6	1	0.788	0.462
hit the road	7	1	6	0.462	0.462
cut corners	10	1	9	0.091	0.474
turn the tables	15	1	14	0.489	0.483
under the table	23	23	0	0.115	0.500
against the grain	17	0	17	0.500	0.500
run out of steam	16	0	16	0.059	0.500
rock the boat	11	0	11	0.154	0.500
flash in the pan	5	0	5	0.500	0.500
in broad daylight	5	0	5	0.000	0.500
mend fences	5	0	5	0.500	0.500
all wet	4	4	0	0.000	0.500
drop the ball	4	4	0	0.500	0.500
miss the boat	3	0	3	0.000	0.500
piping hot	3	0	3	0.500	0.500
raise the roof	3	0	3	0.500	0.500
apples and oranges	2	2	0	0.000	0.500
kick the bucket	1	0	1	0.500	0.500
turn the corner	26	20	6	0.422	0.557
tie the knot	6	4	2	0.400	0.667
watering hole	6	1	5	0.143	1.000
in a rut	5	1	4	0.444	1.000

Table 1: Test Per-idiom Distribution and F1 Scores (Baseline vs Trained).

Key Contributions & Results

- Novelty: First to our knowledge to ground idioms in visual space by fine-tuning MobileCLIP with idiom literal & figurative images.
- Performance: Accuracy improved from 0.51 \rightarrow 0.90, weighted F1 \rightarrow 0.90.
- Techniques driving gains:
 - Span-to-image alignment
 - Pairwise margin & Supervised contrastive losses
 - Auxiliary COCO retrieval as regularizer (prevented overfitting, even improved COCO T2I@1).

Challenges & Limitations

- Literal spans weaker: Model mistakes bias towards figurative.
- Ambiguity & imbalance: Certain idioms (few examples, single-sense) inflate metrics—highlights need for per-idiom evaluation.
- Noise in image anchors: Generated images sometimes failed to represent idiomatic nuance.
- Small test sizes: Some idioms had as few as 1–5 examples.

Future Directions

- Modeling improvements:
 - Hyperparameter optimization (e.g., Optuna)
 - Alternative losses (triplet, angular margin, center loss)
 - Richer visual features & data augmentation (esp. for literal spans).
- Evaluation: Balance and increase idiom datasets.
- Extensions: Apply framework to metaphors, sarcasm, and broader figurative language tasks