Semantic Contrastive Adaptation for Multimodal Figurative Language Understanding

Ayaan Siddiqui
BITS Pilani Hyderabad Campus
f20231060@hyderabad.bits-pilani.ac.in

Abstract

Understanding idiomatic and figurative language in images remains a fundamental challenge for vision–language models, as it requires reasoning beyond literal image–text alignment. Although large pretrained models such as CLIP and BLIP-2 perform well on literal recognition, they consistently fail on multimodal figurative benchmarks, often favoring visually salient but semantically literal interpretations. We show that this failure arises from a systematic literal alignment bias rather than limited model capacity. Motivated by this observation, we reformulate multimodal figurative understanding as a contrastive semantic deviation problem, where figurative images must be distinguished from visually plausible literal alternatives. We introduce a parameter-efficient adaptation of CLIP using Low-Rank Adaptation (LoRA) with hard literal negative mining, achieving targeted reshaping of multimodal representations without full fine-tuning. Experiments on the IRFL benchmark across idioms, metaphors, and similes demonstrate substantial improvements over zero-shot CLIP, BLIP-2, ensemble-based, and knowledge-augmented baselines. Finally, we introduce FIGMENT, a multilingual figurative grounding evaluation spanning five idiom-rich languages, and show that the adapted model generalizes across languages despite being trained exclusively on English supervision.

1 Introduction

Figurative language including idioms, metaphors, and similes conveys meaning that cannot be derived from literal word interpretation alone and plays a central role in human communication and language understanding (Fass, 1991; Birke and Sarkar, 2006). In contemporary visual media such as memes, advertisements, and social platforms, figurative meaning is increasingly expressed through text–image combinations, requiring models to reason jointly over visual and linguistic signals. This has given rise to growing interest in multimodal figurative language understanding, where systems must go beyond surface-level alignment to capture abstract and contextual meaning.

Recent vision–language models such as CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023) align image and text representations through large-scale contrastive pretraining and achieve strong zero-shot generalization on a wide range of literal tasks. However, multiple studies show that these models struggle with figurative language. For example, CLIP achieves only 22% accuracy on the IRFL benchmark (Yosef et al., 2023), whereas human performance on the same task is near perfect. We observe similarly low zero-shot performance for BLIP-2 on figurative idioms, suggesting that increased model capacity alone does not resolve this challenge.

A key reason for this failure lies in how figurative meaning is represented. Figurative meaning is not an intrinsic property of an image–text pair, but instead emerges relative to plausible literal interpretations. In datasets such as IRFL (Yosef et al., 2023), models are required to distinguish figurative images from visually compelling literal alternatives that strongly align with surface semantics. Similarity-based scoring and binary classification approaches therefore tend to favor literal interpretations, even when they are semantically incorrect (Li et al., 2024).

This observation motivates the perspective that rather than treating figurative understanding as absolute classification, it should be modeled as a problem of semantic deviation from literal grounding. Under this view, successful figurative reasoning requires explicitly suppressing strong literal confounders in favor of abstract, non-literal interpretations. In the following sections, we formalize this intuition and show how it can be operationalized through contrastive training with hard literal negatives and parameter-efficient adaptation.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 142–151

(a) (b)

Figure 1: Literal and figurative visual interpretations of the idiom "break the ice". The figurative image depicts social interaction and reduced interpersonal tension, while the literal image depicts physical ice being broken. In our evaluation, models must associate the idiom with the figurative image rather than the literal distractor.

2 Related Work

2.1 Multimodal Figurative Language

Recent work has begun to explore figurative language in multimodal contexts, extending beyond text-only analysis to settings where meaning emerges from the interaction of visual and linguistic cues. The IRFL benchmark (Yosef et al., 2023) is a key evaluation dataset for multimodal figurative language, containing English idioms, metaphors, and similes paired with both figurative and literal images. In this benchmark, models are tasked with selecting which image best represents the figurative meaning of a given phrase, requiring discrimination between visually plausible literal interpretations and the intended figurative meaning. Datasets such as V-FLUTE (Saakyan et al., 2025) provide similar multimodal figurative evaluation, alongside related efforts in multimodal metaphor, humor, and meme understanding (Hasan et al., 2019). Baseline approaches on these benchmarks typically rely on similarity-based scoring with pretrained vision–language models or frame figurative understanding as a binary classification problem. While these methods provide a useful starting point, they consistently exhibit large gaps to human performance, reflecting broader limitations of current vision–language models in abstract and non-literal reasoning.

2.2 Vision–Language Models

Large pretrained vision–language models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP-2 (Li et al., 2023) learn joint representations through contrastive pretraining on large-scale image–text data. These models demonstrate strong zero-shot generalization on a wide range of recognition and retrieval tasks. However, several studies have shown that they struggle with abstract reasoning, non-literal language, and culturally grounded semantics. Our findings align with this line of work, showing that even advanced models such as BLIP-2 fail on multimodal figurative benchmarks, indicating that the challenge lies beyond model capacity.

2.3 Parameter-Efficient Adaptation

Parameter-efficient fine-tuning methods, including adapters, prefix tuning, and Low-Rank Adaptation (LoRA), have emerged as effective techniques for specializing large pretrained models with minimal additional parameters (Hu et al., 2022; Houlsby et al., 2019; Li and Liang, 2021). LoRA has been successfully applied to both language models and vision–language architectures, enabling task adaptation without full fine-tuning. Prior work has primarily used such methods for efficiency or domain adaptation. In contrast, we leverage LoRA as a mechanism to reshape multimodal representations under a *contrastive semantic deviation objective*, rather than as a standalone solution.

2.4 Contrastive Learning with Hard Negative Mining

Hard negative mining has long been used in metric learning and contrastive representation learning to improve discriminative power by focusing on confusable examples (Schroff et al., 2015). In vision–language settings, contrastive objectives typically sample negatives randomly or within batches (Chen et al., 2020). Recent work has shown that incorporating carefully selected hard negatives can substantially improve multimodal representation learning (Radenovic et al., 2023). We build on this idea by explicitly mining *hard literal negatives*—literal images that strongly align with figurative phrases under pretrained models—to model semantic deviation. Unlike prior work, our approach targets the specific failure mode of literal alignment bias in multimodal figurative understanding.

3 Methodology

We develop our approach by first analyzing why pretrained vision–language models fail at multimodal figurative understanding, and then introducing a targeted adaptation strategy to address this failure mode. In particular, we show that methods which operate primarily through representation aggregation or textual enrichment largely preserve a strong bias toward literal image–text alignment

Categories	Idiom	Metaphor	Simile	
			Cl.	Op.
CLIP-ViT-B/16	17	25	52	40
CLIP-ViT-B/32	16	23	45	38
CLIP-RN50	14	27	47	35
CLIP-RN50x64	**22**	**30**	52	**41**
BLIP2	19	19	**57**	40

Table 1: Zero-shot accuracy (%) on the IRFL figurative language benchmark across three figurative categories: idioms, metaphors, and similes (both closed-form Cl. and open-form Op. variants). Values show the performance of different CLIP architectures and BLIP-2 without any adaptation.

(Yosef et al., 2023; Li et al., 2023). Motivated by this analysis, we reformulate figurative understanding as a problem of *semantic deviation from literal grounding* and introduce a parameter-efficient contrastive adaptation strategy based on hard literal negatives.

3.1 Ensemble-Based CLIP Baselines

A natural baseline for improving figurative image understanding is to leverage architectural diversity through ensembling pretrained CLIP variants. Let $\mathbf{v}_i \in R^d$ and $\mathbf{t}_i \in R^d$ denote the image and text embeddings produced by the i-th CLIP variant, where $i \in \{1, \ldots, 4\}$ indexes ViT-B/32, ViT-B/16, RN50, and RN50×64. For each variant, cosine similarity is computed as:

$$s_i(x, y) = \frac{\mathbf{v}_i(x) \cdot \mathbf{t}_i(y)}{\|\mathbf{v}_i(x)\| \|\mathbf{t}_i(y)\|}. \quad (1)$$

Weighted Similarity Averaging. We combine similarity scores via a convex combination:

$$S_{\text{ens}}(x, y) = \sum_{i=1}^{4} w_i \, s_i(x, y), \quad \sum_{i=1}^{4} w_i = 1, \; w_i \geq 0, \quad (2)$$

Ensemble Weights				Idiom Detection	
B/32	B/16	RN50	50x64	Phrase	+ Def
0.1	0.1	0.2	0.6	18	19
0.1	0.2	0.3	0.4	17	17
0.2	0.1	0.4	0.3	16	16
0.5	0.1	0.2	0.2	18	18

Table 2: Sampled ensemble weight configurations and their accuracy (%) on IRFL idiom detection. Multiple weight combinations were tested on validation data; the presented rows show representative configurations. Abbreviations: B/32 = CLIP-ViT-B/32, B/16 = CLIP-ViT-B/16, RN50 = CLIP-RN50, 50x64 = CLIP-RN50x64.

Model	Idioms	Metaphors	Similes
CLIP (IRFL Zero Shot)	22	**30**	44
Ours: Weighted Ens.	22	25	44
Ours: Supervised MLP	48	–	–
Ours: LoRA-CLIP	**72**	**48**	**62**
Human (IRFL)	97	99.7	100

Table 3: Performance of methods on the IRFL "mixed" multimodal figurative language detection task. Human scores are reported from the IRFL paper.

where the weights w_i are tuned on validation data (Dietterich, 2000). This approach aims to exploit complementary inductive biases across CLIP backbones.

While ensembling yields modest improvements over single-model baselines, we observe that it largely preserves the same failure patterns. In particular, images that strongly align with the surface semantics of a phrase continue to dominate predictions, even when they correspond to literal rather than figurative interpretations. This indicates that aggregation alone cannot overcome the literal alignment bias inherent in pretrained embeddings.

Learned Fusion via MLP. To allow for nonlinear interactions between models, we also explore learned fusion using a multilayer perceptron (MLP). Image and text embeddings from all CLIP variants are concatenated,

$$\boldsymbol{h} = [\boldsymbol{v}_1, \boldsymbol{t}_1, \ldots, \boldsymbol{v}_4, \boldsymbol{t}_4] \in R^{8d}, \quad (3)$$

and mapped to a scalar score through a three-layer feedforward network. Despite this increased flexibility, empirical results show that learned fusion does not substantially outperform simpler ensemble averaging, suggesting that nonlinear aggregation alone is insufficient to induce figurative abstraction.

spill the beans
(spill the beans, `Synonym`, confess)
(spill, `RelatedTo`, reveal) (beans, `UsedFor`, food)

cold feet
(cold feet, `RelatedTo`, nervous)
(cold, `Antonym`, hot) (feet, `PartOf`, body)

break the ice
(break the ice, `Synonym`, introduce)
(break, `Causes`, pieces) (ice, `RelatedTo`, cold)

Table 4: Examples of ConceptNet augmentation for idiomatic expressions. Figurative relations were concatenated in the text encoder to enrich embeddings.

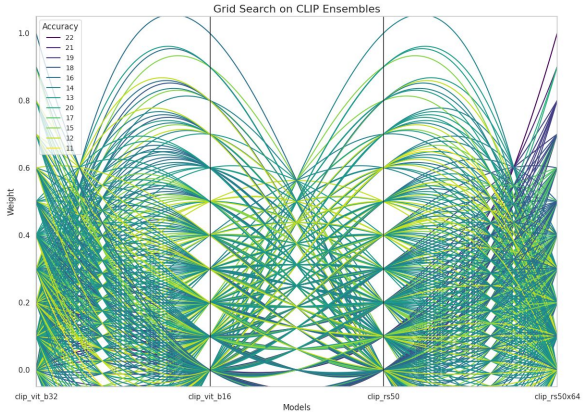

Figure 2: Grid search over CLIP ensemble weight configurations. The parallel-coordinates plot shows accuracy across different weight combinations, where each line represents a different weighting of the four CLIP variants (ViT-B/32, ViT-B/16, RN50, RN50×64). No consistent improvement is achieved over the strongest single CLIP backbone (RN50×64), highlighting the limitations of representation aggregation for figurative understanding.

3.2 ConceptNet Knowledge Augmentation

Another commonly adopted strategy for figurative understanding is to enrich textual representations with external commonsense knowledge. We augment idiomatic definitions using ConceptNet (Speer et al., 2017), with phrases as demonstrated by examples in Table 4. For a phrase definition y, we retrieve a set of related concepts $\mathcal{C}(y) = \{c_1, \ldots, c_k\}$ and construct an augmented input:

$$y' = y \oplus \{c_1, \ldots, c_k\}, \tag{4}$$

where $\oplus$ denotes concatenation. The resulting text embedding $\mathbf{t}_i(y')$ is then used in ensemble scoring.

Although ConceptNet augmentation occasionally improves performance for certain idioms, its effects are inconsistent across figurative categories. Crucially, enriching text alone does not modify the visual representation space: visually salient literal images continue to receive high similarity scores. This further suggests that multimodal figurative understanding requires altering how visual and textual representations interact, rather than adding external knowledge at the text level.

3.3 Contrastive Semantic Deviation with LoRA

The limitations of ensemble fusion and knowledge augmentation motivate a more direct intervention: explicitly reshaping the joint image–text embedding space to distinguish figurative interpretations from visually plausible literal alternatives. We therefore formalize multimodal figurative understanding as a *contrastive semantic deviation* problem.

Let $f_\theta^I(\cdot)$ and $f_\theta^T(\cdot)$ denote the image and text encoders of a pretrained vision–language model with parameters θ. Given a figurative phrase p and an image I, the model assigns a similarity score

$$s(p, I) = \cos\left(f_\theta^T(p), f_\theta^I(I)\right). \tag{5}$$

In pretrained models, literal images I^{lit} often achieve higher similarity than the intended figurative image I^{fig}, yielding

$$s(p, I^{\text{lit}}) \geq s(p, I^{\text{fig}}), \tag{6}$$

which we identify as *literal alignment bias*.

We seek to learn a parameter update $\Delta\theta$ such that figurative images are ranked above their literal counterparts while preserving general visual–semantic alignment. To this end, we define a contrastive objective over triples $(p, I^{\text{fig}}, I^{\text{lit}})$:

$$\mathcal{L}_{\text{dev}} = -\log \frac{e^{s(p, I^{\text{fig}})/\tau}}{e^{s(p, I^{\text{fig}})/\tau} + e^{s(p, I^{\text{lit}})/\tau}}, \tag{7}$$

which can be equivalently formulated as maximizing the margin $m = s(p, I^{\text{fig}}) - s(p, I^{\text{lit}})$, where τ is a temperature parameter. This objective explicitly enforces semantic deviation by increasing the similarity difference between figurative and literal images.

To efficiently optimize this objective without full fine-tuning, we apply Low-Rank Adaptation (LoRA) to a frozen CLIP backbone. For a pretrained weight matrix $W_0 \in R^{d\times k}$, LoRA parameterizes the update as

$$W = W_0 + \alpha AB, \tag{8}$$

where $A \in R^{d\times r}$, $B \in R^{r\times k}$, and $r \ll \min(d, k)$ (i.e., the rank r is much smaller than both the input and output dimensions). During training, only A and B are optimized, while W_0 remains fixed. This constrains learning to a low-dimensional subspace, enabling targeted reshaping of the embedding geometry rather than wholesale reparameterization.

We pair LoRA adaptation with *hard literal negative mining* (Radenovic et al., 2023). For each phrase p, we identify literal images I^{lit} that maximize $s(p, I)$ under the frozen base model and use them as negatives during training. This focuses learning on confusable literal cases that expose representational bias, rather than easy or randomly sampled negatives.

Task	Single CLIP	Simple	Weighted	Max-Sim
Idioms	22	22	22	21
Metaphors	23	25	25	**27**
Similes	**45**	44	44	41

Table 5: Accuracy (%) of CLIP ensemble variants with ConceptNet augmentation across figurative categories.

4 Experimental Setup

4.1 Dataset and Task

We evaluate all methods on the IRFL dataset (Yosef et al., 2023), which contains English idioms, metaphors, and similes paired with both figurative and literal images. Given a figurative phrase and a set of candidate images, the task is to select the image that best represents the figurative meaning of the phrase. This formulation naturally reflects the relative nature of figurative understanding, where models must prefer figurative interpretations over visually plausible literal alternatives.

We follow the IRFL train, validation, and test splits provided with the dataset and report results separately for idioms, metaphors, and similes.

4.2 Baselines

We compare our approach against several baselines designed to probe the limitations of pretrained vision–language models under figurative reasoning:

Zero-shot CLIP. We evaluate CLIP in a zero-shot setting using cosine similarity between image and text embeddings.

Ensemble-Based CLIP. We evaluate multiple CLIP variants using simple averaging, weighted similarity averaging, and max-similarity aggregation. These baselines test whether architectural diversity alone can mitigate literal alignment bias.

ConceptNet-Augmented CLIP. We additionally augment phrase definitions with external commonsense knowledge from ConceptNet and evaluate ensemble-based variants using the augmented text representations. These baselines test whether enriching textual semantics is sufficient for multimodal figurative understanding.

Across these baselines, performance typically ranges between 14–27% accuracy, consistent with prior findings on the difficulty of multimodal figurative language understanding (Saakyan et al., 2025).

4.3 Training Details

For all CLIP-based methods, we use publicly available pretrained checkpoints and ℓ_2-normalize image and text embeddings prior to similarity computation. Ensemble weights are tuned on validation data. Learned fusion models are trained using Adam optimizer with a learning rate of 10^{-4} and early stopping.

For LoRA-based adaptation, we fine-tune the ViT-B/32 CLIP backbone for 10 epochs using a learning rate of 10^{-5}. We implement LoRA using the PEFT library (Mangrulkar et al., 2022), keeping all base CLIP parameters frozen.

To analyze where parameter-efficient adaptation is most effective, we evaluate LoRA applied to different CLIP subcomponents and layers. Table 6 reports accuracy across idioms, metaphors, and similes for various configurations.

5 Results

5.1 Ensemble Strategies

We first examine whether architectural diversity improves multimodal figurative understanding. Simple aggregation of similarity scores across CLIP variants yields modest robustness gains, consistent with standard ensemble learning approaches (Dietterich, 2000). However, grid search over ensemble weights (Figure 2) shows that re-weighting CLIP backbones does not surpass the strongest individual model (RN50×64).

Across idioms, ensemble accuracy remains in the 16–19% range, regardless of whether phrase definitions are included. Learned fusion using multilayer perceptrons similarly provides limited benefit. These findings indicate that architectural diversity alone largely amplifies existing literal alignment biases rather than inducing figurative abstraction. While ensembling improves robustness for literal recognition tasks, it does not substitute for targeted modification of multimodal representations.

5.2 ConceptNet-Based Knowledge Augmentation

Table 5 reports performance for CLIP backbones and ensemble variants augmented with external commonsense knowledge from ConceptNet. Zero-shot CLIP achieves only 22% accuracy on idioms and 23% on metaphors, consistent with prior observations that figurative meaning is poorly captured in pretrained multimodal embeddings (Yosef et al., 2023).

Configuration	Idioms		Metaphors	Similes	
	Phr	+ Def		Cl.	Op.
text_projection + vision_projection	69	69	31	64	53
LoRA applied on layer $i = 11$					
text_attention only	63	65	33	64	56
vision_attention only	70	**71**	38	66	60
text_attention + text_projection	67	66	40	66	59
vision_attention + vision_projection	**71**	70	42	61	53
LoRA applied on layers $i \in \{9, 10, 11\}$					
text_attention only	68	69	31	63	56
vision_attention only	70	**71**	45	**68**	**62**
text_attention + text_projection	69	70	42	65	58
vision_attention + vision_projection	68	68	46	63	55
(text + vision)_attention + (text + vision)_projection	**72**	**72**	**48**	64	60

Table 6: LoRA-adapted fine-tuning accuracy (%) on the IRFL dataset when applied to different CLIP ViT-B/32 subcomponents. Rows show different LoRA configurations: "Projections only" adapts only the projection layers; "Text/Vision attention only" adapts attention heads; "Text + text_projection" and "Vision + vision_projection" adapt attention and projection layers together. Columns report performance across idioms, metaphors (Metaphors), and simile subtypes (Closed Cl. and Open Op.). The best configurations achieve substantial gains when LoRA is applied to vision attention layers, especially across layers $\{9, 10, 11\}$.

Augmenting phrase definitions with ConceptNet relations yields small but consistent gains, with max-similarity ensembles reaching up to 27% accuracy on metaphors. However, improvements on idioms and similes remain marginal. These results suggest that while external knowledge can enrich textual semantics, it does not address the core challenge of multimodal figurative understanding: visually salient literal interpretations continue to dominate similarity-based scoring. Enriching text alone is therefore insufficient without modifying how visual representations are aligned.

5.3 LoRA Fine-Tuning and Contrastive Adaptation

Table 6 shows that LoRA-based adaptation under the proposed contrastive semantic deviation objective substantially outperforms both ensemble methods and knowledge augmentation across all figurative categories. The best configurations achieve up to 72% accuracy on idioms, representing a large absolute improvement over zero-shot CLIP and indicating a successful reversal of literal alignment bias.

Across categories, vision-side adaptation consistently yields larger gains than text-side adaptation, particularly when LoRA is applied to attention layers in higher transformer blocks. This pattern suggests that figurative grounding depends critically on reshaping abstract visual representations and their interaction with language, rather than modifying surface-level textual semantics. In terms of the deviation objective, vision-layer adaptation most effectively increases the similarity margin $s(p, I^{\text{fig}}) - s(p, I^{\text{lit}})$, directly enforcing the desired ranking between figurative and literal images.

Text-side adaptation benefits idiomatic expressions when phrase definitions are included, indicating that explicit semantic grounding can assist language representations when paired with contrastive deviation training. However, these effects are weaker and less consistent for metaphors and similes, where visual context and relational structure play a more dominant role. This asymmetry highlights the central role of visual abstraction in multimodal non-literal reasoning.

For metaphors, LoRA improves performance to approximately 48% accuracy, while simile understanding benefits most from multi-layer vision adaptation, reaching 68% accuracy in closed settings and 62% in open settings. Despite these gains, a substantial gap remains to human-level performance on idioms (97%), indicating that figurative understanding remains a challenging open problem even under targeted adaptation.

Overall, these results provide empirical support for our contrastive semantic deviation formulation: explicitly reshaping high-level vision–language representations is necessary to suppress literal alignment and promote figurative interpretation. Parameter-efficient adaptation via LoRA enables this reshaping in a controlled and computationally lightweight manner.

Language	Idioms	Images	% of Dataset
Hindi	70	140	28.1
Arabic	47	94	18.9
French	49	98	19.7
Chinese	44	88	17.7
Spanish	39	78	15.7
Total	**249**	**498**	**100.0**

Table 7: Language-wise distribution of idioms and images in FIGMENT. Each idiom is paired with one literal and one figurative image.

6 FIGMENT: Figurative Idiom Grounding in Multilingual Environments

We introduce FIGMENT, a multilingual evaluation framework for multimodal figurative grounding spanning five idiom-rich languages: Hindi, French, Chinese, Spanish, and Arabic. Using FIGMENT, we evaluate figurative grounding under systematic cross-lingual distribution shift, examining whether the literal alignment bias observed in English benchmarks persists across languages and whether semantic deviation learned from English supervision transfers to linguistically and culturally diverse contexts. Table 7 summarizes the language distribution.

6.1 Dataset Construction

FIGMENT is constructed using a controlled, semi-automated multi-agent pipeline designed to obtain visually grounded literal and figurative representations of idiomatic expressions. For each idiom, an initial agent generates a concise English gloss capturing the idiomatic meaning. A second agent converts this representation into two concrete, image-searchable prompts corresponding to the literal and figurative interpretations.

Images are retrieved from public web image search APIs using only the top-ranked result per prompt to simulate realistic, weakly supervised data collection (Schuhmann et al., 2022; Radford et al., 2021). To ensure visual grounding quality, a final validation stage filters out text-heavy or meme-style images using OCR-based text detection (Li et al., 2020), and rejected images are re-queried with disambiguated prompts. Approximately 15% of images required manual replacement due to persistent retrieval or validation failures. All remaining examples are manually verified for correctness and visual distinctness. FIGMENT is not used for training and is intended solely for diagnostic evaluation of figurative grounding behavior.

6.2 Models and Metrics

We compare three classes of models: (i) zero-shot multilingual CLIP (mCLIP; Chen et al., 2023), (ii) zero-shot BLIP-2, and (iii) our LoRA-adapted model trained exclusively on English IRFL data. All models are evaluated in a binary forced-choice setting, where the objective is to rank the figurative image above a visually plausible literal alternative.

In addition to accuracy (Precision@1), we report two diagnostic metrics that capture figurative grounding behavior. We define the similarity margin as $s(p, I^{\text{fig}}) - s(p, I^{\text{lit}})$, measuring the difference between figurative and literal image similarity scores, with positive values indicating successful semantic deviation, following standard contrastive learning formulations (van den Oord et al., 2018). The *literal bias rate* measures the percentage of cases in which the literal image is ranked higher than the figurative image, directly reflecting a model's tendency to favor literal interpretations.

6.3 Quantitative Results Overview

Table 8 summarizes multilingual figurative grounding performance across all five languages. Zero-shot mCLIP performs far below the random baseline (50%), with accuracies ranging from 4% to 28%. This failure is accompanied by strongly negative similarity margins and extreme literal bias (up to 96%), indicating a systematic preference for visually plausible literal interpretations despite multilingual pretraining.

BLIP-2 exhibits modest improvements over mCLIP in some languages, but remains near chance overall. Similarity margins are clustered around zero, and literal bias remains substantial. These results suggest that generative vision–language models do not inherently resolve multimodal figurative grounding when deployed in zero-shot retrieval settings.

6.4 Effect of Contrastive Adaptation

In contrast, our LoRA-adapted model achieves consistent gains across all five languages, with accuracies ranging from 68% to 85%. Importantly, similarity margins are uniformly positive and literal bias is reduced by more than two-fold compared to both mCLIP and BLIP-2. Since the model is trained exclusively on English idioms, these results indicate that contrastive semantic deviation operates on language-agnostic representational structure rather

Language	Model	Acc (%)	Margin	LitBias (%)
French	mCLIP	4.1	−0.060	95.9
	BLIP-2	46.9	−0.011	53.1
	Ours	**73.5**	+1.564	**26.5**
Chinese	mCLIP	20.5	−0.040	79.5
	BLIP-2	36.4	−0.043	63.6
	Ours	**72.7**	+1.143	**27.3**
Spanish	mCLIP	18.0	−0.047	82.1
	BLIP-2	43.6	−0.024	56.4
	Ours	**84.6**	+2.047	**15.4**
Hindi	mCLIP	28.8	−0.028	71.2
	BLIP-2	53.7	+0.000	46.3
	Ours	**68.7**	+1.286	**31.3**
Arabic	mCLIP	23.4	−0.040	76.6
	BLIP-2	57.5	+0.001	42.6
	Ours	**72.3**	+1.345	**27.7**

Table 8: Accuracy, similarity margin, and literal alignment bias across five languages of FIGMENT dataset.

than surface-level lexical alignment (Feng et al., 2022).

6.5 Margin–Bias Relationship and Cross-Lingual Generalization

Figure 3 provides quantitative evidence that the similarity margin–bias relationship underlies robust figurative grounding. Zero-shot CLIP variants exhibit near-zero or negative margins (-0.09 to -0.03), corresponding to substantial literal bias (53–96%), whereas the LoRA-adapted model achieves positive margins (avg. $+1.14$) with substantially reduced bias (15–31%).

This margin–bias relationship holds consistently across all five languages (French, Chinese, Spanish, Hindi, and Arabic), demonstrating strong cross-lingual generalization and highlighting explicit suppression of literal bias as central to figurative grounding.

6.6 Accuracy Across Languages

This section compares figurative grounding accuracy across languages for all three models. While mCLIP and BLIP-2 fluctuate around chance performance, our LoRA-adapted model consistently achieves high accuracy across all five languages, demonstrating robust cross-lingual generalization.

Figure 4 provides a mechanistic explanation for these accuracy trends by relating figurative grounding performance to the similarity margin between figurative and literal images. Across all five languages, zero-shot CLIP variants cluster around zero or negative margins, corresponding to strong literal alignment bias and low accuracy. BLIP-2 shows

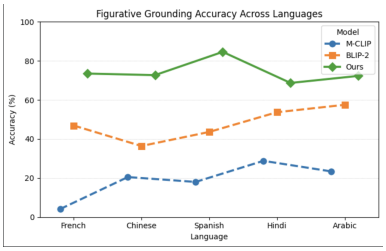

Figure 3: Figurative grounding accuracy across five languages on FIGMENT for zero-shot mCLIP, zero-shot BLIP-2, and our LoRA-adapted model.

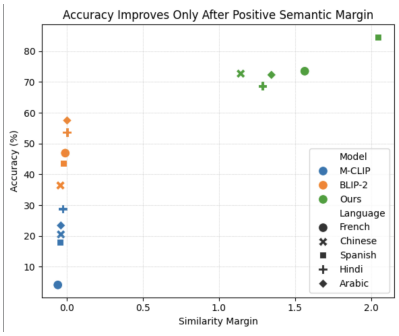

Figure 4: Similarity margin vs. literal alignment bias across CLIP variants and five languages. Negative or near-zero margins correspond to strong literal bias, while positive margins achieved by our LoRA-adapted model reduce literal bias and improve figurative grounding.

slightly reduced bias but remains near zero margin, explaining its near-chance performance. In contrast, our LoRA-adapted model consistently achieves positive semantic margins, explicitly separating figurative images from literal alternatives in the embedding space. This shift is accompanied by substantially higher accuracy across languages, indicating that robust figurative grounding emerges only once literal alignment is actively suppressed.

7 Conclusion

This work studies multimodal figurative language understanding through the lens of *semantic deviation*, highlighting the difficulty of separating figurative meanings from visually plausible literal interpretations. We show that pretrained vision–language models strongly favor literal image–text alignment, and that common techniques such as ensembling and knowledge augmentation yield only limited gains, as they largely preserve the original representations.

In contrast, parameter-efficient adaptation with LoRA, combined with contrastive training against hard literal negatives, substantially improves fig-

urative grounding across idioms, metaphors, and similes. Gains are strongest when adapting higher-level visual attention layers, suggesting that figurative meaning relies on abstract visual representations. Text-side adaptation helps for idioms when definitions are available, but visual adaptation is more consistently effective. Evaluations on FIGMENT, a multilingual benchmark introduced here, show that semantic deviation learned from English generalizes across multiple languages. While a gap to human performance remains, our results demonstrate that targeted, lightweight adaptation can meaningfully advance multimodal non-literal understanding.

Limitations

While the proposed approach shows consistent improvements in multimodal figurative grounding, these limitations highlight promising directions for future work.

First, although contrastive semantic deviation effectively reduces literal alignment bias, performance on idioms remains below human-level accuracy. This suggests that fully capturing figurative meaning may benefit from richer contextual, cultural, or pragmatic cues, which are not explicitly modeled in current vision–language representations (Alhamoud et al., 2025).

Second, our method relies on supervised figurative datasets in which figurative images are paired with visually plausible literal alternatives. While this setup is well suited for diagnosing and correcting the bias, extending the approach to more weakly supervised or open-world settings remains an interesting engineering challenge.

Third, cross-lingual generalization is evaluated on a fixed set of five languages in the FIGMENT benchmark. Although the results indicate encouraging transfer, future work could explore a broader range of languages and figurative expressions, including cases that are more abstract or less visually grounded.

Finally, our experiments focus on CLIP-based architectures adapted using parameter-efficient LoRA fine-tuning. While this choice enables controlled analysis of representational bias, it would be valuable to investigate how the proposed semantic deviation objective interacts with other vision–language architectures or large generative models in future studies (Li et al., 2024).

References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mclip: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17:49–90.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung,

Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Hui Li, Peng Wang, and Chunhua Shen. 2020. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. ICML'23. JMLR.org.

Siting Li, Pang Wei Koh, and Simon Shaolei Du. 2024. On erroneous agreements of CLIP image embeddings. *arXiv preprint arXiv:2411.05195*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision–language pre-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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 (NAACL)*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: an open large-scale dataset for training next generation image-text models.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.