Beyond Literal Bias: Association-Guided Evidence Search for Multimodal Figurative Language Detection

Anonymous ACL submission

Abstract

Multimodal Figurative Language Detection aims to map textual figurative expressions to their visual counterparts. Formulated as a cross-modal, multi-instance matching task under the challenge of carefully designed literal distractors, it rigorously evaluates a model's capacity for abstract reasoning. While prior research indicates that standard VL-PTMs are vulnerable in this setting, the potential of Multimodal Large Language Models (MLLMs) remains largely unexamined. In this work, we reveal that even advanced MLLMs fall short. We identify two primary bottlenecks for this failure: visual symbolic misinterpretation, where models fail to recognize the abstract symbolism in an image, and literal bias, where models correctly understand the symbolism but inadvertently favor images consisting purely of literal elements. To address these issues, we propose AGES, an Association-Guided Evidence Search framework that explicitly models the internal structure of figurative expressions and leverages an association mechanism to guide attention toward visual evidence that supports the figurative ground. Experimental results show that AGES effectively suppresses literal distractors and improves figurative image-text alignment, advancing multimodal figurative language understanding.

1 Introduction

Figurative language refers to the use of words or expressions that convey meanings beyond their literal interpretation to achieve a more complex or powerful effect. As a fundamental component of human interaction, it allows for the expression of abstract concepts, subtle emotions, and implicit intentions that literal lexicon often fails to capture. For Artificial Intelligence, the transition from decoding what is explicitly stated to inferring what is truly meant remains a formidable challenge.

Recently, multimodal figurative language understanding has gained significant traction, driven by the ubiquity of figurative imagery in memes and digital advertisements. While extensive research has explored figurative language recognition or interpretation on an isolated, single-instance level (Zhang et al., 2021; Xu et al., 2022; Liu et al., 2022; Zhang et al., 2023; Hwang and Shwartz, 2023; Saakyan et al., 2024), the ability of models to perform cross-modal, multi-instance matching remains largely overlooked. To investigate this challenge, Yosef et al. (2023) introduced the task of **Multimodal Figurative Language Detection**. Crucially, they formulated this detection paradigm as a multiple-choice matching problem: given a figurative text query and four candidate images, the model must select the single image that best conveys the figurative meaning. A defining characteristic of this task is the deliberate inclusion of **literal distractors**–images that depict the literal constituents of a query but fail to convey its underlying figurative essence. Initial evaluations indicated that standard Vision-Language Pre-trained Models (VL-PTMs), such as CLIP, perform poorly in these settings. This failure stems from their reliance on shared embedding spaces that prioritize superficial semantic correspondence. Thus, when presented with a query like 'heart of gold,' these models gravitate toward images of anatomical hearts or gold bars rather than depictions of kindness or generosity.

Although Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have shown promise in interpreting figurative language, our preliminary experiments reveal a surprising finding: even advanced MLLMs remain strongly affected by literal distractors. As shown in Figure 1 (a), they suffer a significant performance degradation when literal distractors are introduced. To investigate the source of distractor effects, we evaluated models under two settings. In the first setting, the model analyzed each candidate image independently, assigned a matching score based on

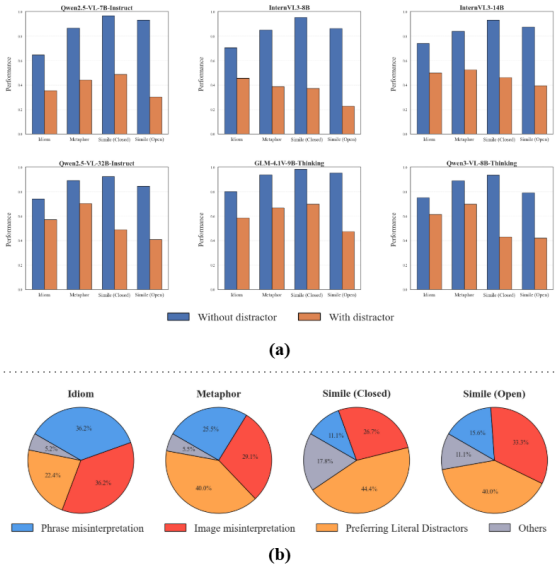


Figure 1: Our preliminary studies: (a) Performance drop of various MLLMs on multimodal figurative language detection when introducing literal distractors. (b) Error distribution across four figurative language types.

its alignment with the figurative expression, and selected the highest-scoring image as the answer. In the second setting, all candidate images were presented simultaneously, allowing the model to choose the relatively best option through comparison. The first setting performed slightly worse overall, suggesting that independent evaluation does not eliminate distractor effects. Instead, the tendency to favor literal images appears to be an inherent model bias, while separate scoring may further reduce performance because the correct choice is often comparative rather than absolute. Based on these findings, we conducted a more detailed error analysis under the second setting (Figure 1 (b)) and identified two major error patterns: (1) difficulty in correctly interpreting how the target image symbolically reflects the intended figurative meaning, and (2) a persistent preference for literal images even when the figurative meaning is partially understood. While some errors in idiomatic expressions arise from a lack of world knowledge, we find that simply augmenting knowledge provides marginal gains and fails to address these fundamental alignment issues.

To bridge this gap, we propose **A**ssociation-**G**uided **E**vidence **S**earch, a training-free framework for Multimodal Figurative Language Detection. AGES explicitly deconstructs the internal structure of figurative expressions–identifying the *tenor*, *vehicle*, and *ground*–and employs an association mechanism to guide the model in searching for key visual evidence that supports the ground. By shifting the focus from literal overlaps to metaphorical cues within specific image regions, AGES enables robust reasoning even amidst distracting literal information. Extensive experiments demonstrate that AGES substantially outperforms existing baselines in Multimodal Figurative Language Detection, marking a significant step toward deep multimodal figurative understanding.

Our main contributions are summarized as follows:

- We provide a systematic evaluation of representative open-source MLLMs on multimodal figurative language detection, uncovering a persistent bias toward literal visual interpretations.
- We propose **AGES** (Association-Guided Evidence Search), a training-free framework that explicitly decomposes figurative expressions into *tenor*, *vehicle*, and *ground*, and guides models to search for visual evidence supporting the figurative ground.
- Extensive experiments across diverse foundation models show that AGES effectively suppresses literal distractors by emphasizing semantically grounded visual evidence rather than superficial literal overlaps, leading to substantial improvements in multimodal figurative language detection.

2 Related Work

2.1 Figurative Language Understanding

Figurative language involves the use of expressions whose meanings extend beyond their literal interpretation to convey richer or more nuanced effects. Within Natural Language Processing (NLP), metaphor (including simile) has long been the primary focus. Classical linguistic theories decompose metaphors into three structural components–*tenor*, *vehicle*, and *ground*, corresponding to the entity being described, the entity used for comparison, and the shared property underlying the comparison. Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 2024; Lakoff, 1993) further conceptualizes metaphor as a cognitive mapping between domains, where a concrete *source domain* structures the understanding of an abstract *target domain*.

Building on these theoretical foundations, earlier PLM-based methods focused on binary textual

metaphor detection, (Choi et al., 2021; Zhang and Liu, 2023; Li et al., 2023; Tian et al., 2023; Li et al., 2024; Tian et al., 2024). With the emergence of Large Language Models (LLMs), the focus has shifted beyond simple detection toward deeper investigations of metaphor comprehension, including source-domain identification, unsupervised detection, and paraphrase-based evaluation (Chakrabarty et al., 2022; Wachowiak and Gromann, 2023; Yang et al., 2024; Tong et al., 2024; Sengupta et al., 2025). Despite these advances, LLMs still exhibit limitations in complex reasoning and socially-grounded interpretation (Sengupta et al., 2025; Gallipoli and Cagliero, 2025).

Beyond metaphors, idioms (including Chinese Chengyu) represent another significant challenge due to their non-compositional nature (Fu et al., 2025; Qiu et al., 2025). Recent mechanistic analyses suggest that transformers selectively route between literal and figurative representations during inference (Oh et al., 2025). Kim et al. (2025) further indicate that models often rely on a hybrid of memorization and contextual reasoning, rather than a principled semantic representation of idiomatic meaning. While these textual phenomena are extensively studied, most existing literature remains confined to the textual domain, overlooking how figurative expressions are grounded—or potentially misgrounded—in visual contexts.

2.2 Multimodal Figurative Language Understanding

The ubiquity of symbolic imagery in digital media has catalyzed research into multimodal metaphor understanding. Representative benchmarks (Zhang et al., 2021; Xu et al., 2022; Liu et al., 2022; Zhang et al., 2023) provide annotations for metaphor presence, modality, sentiment, intent, and source-target domain mappings, establishing foundational resources for evaluating models' metaphor comprehension capabilities. Subsequent efforts have expanded both the granularity and scope of annotations, such as fine-grained emotion labels (Lu et al., 2025) and cross-cultural variations (Yang et al., 2025). However, these benchmarks predominantly frame the problem as closed-set classification, which limits the assessment of a model's true reasoning capabilities.

To address this limitation, recent work has pivoted toward interpretation-oriented tasks, such as metaphorical captioning, localization, and generation. For instance, Hwang and Shwartz (2023) disentangles literal descriptions from metaphor-aware explanations, encouraging models to distinguish surface content from figurative intent. Akula et al. (2023) further introduces fine-grained visual metaphor annotations with entity relations and bounding boxes, enabling joint recognition, localization, and explanation of metaphorical expressions. Others have explored generating visual content that conveys metaphorical intent (Chakrabarty et al., 2023) or formulated multimodal Natural Language Inference (NLI) tasks to evaluate figurative reasoning (Saakyan et al., 2024).

While preliminary studies on MLLMs have been conducted (Xu et al., 2024; Qian et al., 2025), they focus on single-instance interpretation–analyzing one image-text pair at a time–and fail to explicitly test cross-modal alignment under distraction. In contrast, Yosef et al. (2023) introduces a more stringent evaluation through a multiple-choice setting with literal distractors. This task directly tests whether a model can prioritize figurative alignment over surface-level literal matches. Our work builds upon IRFL but moves beyond simple evaluation by proposing AGES, a framework designed to overcome literal distraction through structural, association-guided evidence searching.

3 Method

3.1 Overview

To mitigate the inherent literal bias in vision-language models and facilitate abstract figurative reasoning, we propose **AGES** (**A**ssociation-**G**uided **E**vidence **S**earch), which transforms figurative image-text matching from an opaque, holistic similarity estimation into a structured, interpretable, and evidence-driven verification process. As illustrated in Figure 2, AGES operates through three synergistic stages: (i) *Figurative Language Decomposition*, (ii) *Component-wise Figurative Verification* anchored by visual perception, and (iii) *Cross-Instance Evidence Comparison*.

3.2 Figurative Language Decomposition

Given a figurative query Q, we first prompt an MLLM to resolve its underlying semantic structure. Following Yosef et al. (2023), we categorize the queries into three distinct figurative types:

- **Idiom**: Phrases whose figurative meaning is highly conventionalized and non-compositional, rendered independent of the literal definitions of their constituent words.

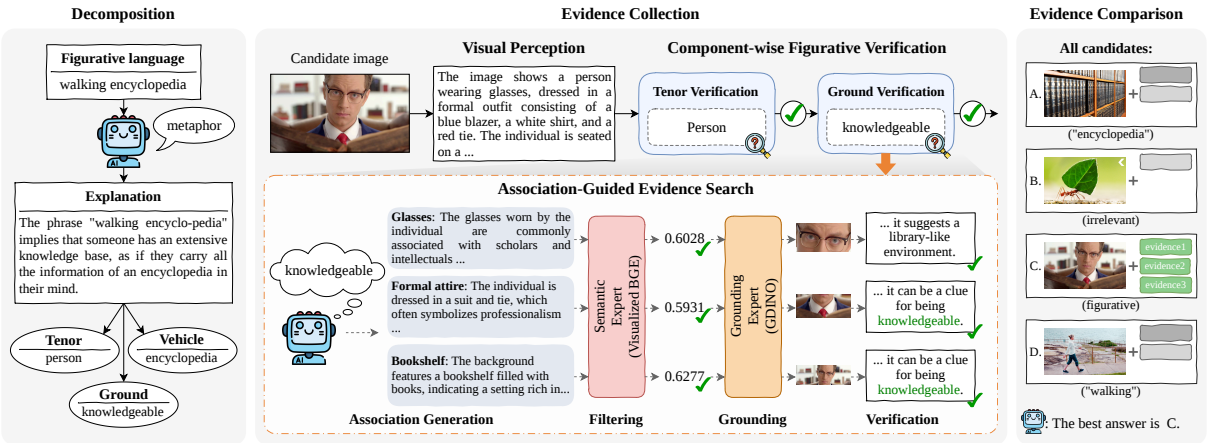


Figure 2: An overview of the AGES framework.

- **Simile**: Explicit comparisons between two dissimilar entities (the *tenor* and the *vehicle*) through the use of "like" or "as". Similes can be further classified as *closed* (explicitly stating the ground, e.g., "A is as ... as B") or *open* (leaving the ground implicit, e.g., "A is like B").
- **Metaphor**: Implicit comparisons between two concepts without explicit linguistic markers such as "like" or "as" (e.g., "A is B"), where the notions of *tenor*, *vehicle*, and *ground* function analogously to those in similes.

Formally, let $\mathcal{P}_{\text{dec}}$ denote the decomposition prompt. We formalize this semantic parsing process to extract the figurative type $t \in \{\text{idiom, simile, metaphor}\}$, the tenor T, the ground G, and a comprehensive natural language explanation E:

$$t, T, G, E = \text{MLLM}(Q, \mathcal{P}_{\text{dec}}) \tag{1}$$

For idioms, which often lack a clear tenor-vehicle duality, we treat their holistic idiomatic meaning directly as the ground G. This structured representation decouples abstract concepts, bypassing the limitations of holistic image-text similarity matching.

3.3 Objective Visual Perception

To provide an objective foundation for subsequent figurative reasoning, we first perform a descriptive analysis of each candidate image I.

Instructed by a neutral description prompt $\mathcal{P}_{\text{desc}}$, the MLLM generates a literal caption C that explicitly captures visible entities, attributes, and spatial relations while strictly avoiding interpretive speculation:

$$C = \text{MLLM}(I, \mathcal{P}_{\text{desc}}) \tag{2}$$

This perception stage serves as a deterministic "visual anchor," ensuring that subsequent high-level figurative judgments are strictly grounded in actual visual evidence.

3.4 Component-wise Figurative Verification

Building upon the decomposed query components and the visual anchor C, we systematically assess the figurative relevance of an image. This process involves verifying a hierarchical duo of conditions: (i) the presence of the tenor (a necessary condition) and (ii) the existence of visual evidence supporting the ground (a sufficient condition).

We first verify whether the candidate image I contains the tenor T. We operationalize this step as a primary semantic filter; if the tenor is absent, the image is deemed irrelevant, thereby effectively eliminating literal distractors that might exclusively depict the vehicle or unrelated salient objects.

If the tenor is confirmed, we evaluate whether the depicted entities satisfy the abstract ground G. This step is notoriously challenging for standard models due to the abstract nature of grounds and their heavy reliance on implicit commonsense reasoning. To bridge this gap, our **Association-Guided Evidence Search** injects commonsense knowledge by a sequence of association generation, semantic filtering, and spatial grounding.

Association Generation. We prompt the MLLM with a specialized prompt $\mathcal{P}_{\text{assoc}}$ to project the abstract ground onto the visual domain, generating a set of candidate visual associations $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ that could plausibly manifest the ground G within the context of image I:

$$\mathcal{A} = \text{MLLM}(I, G, \mathcal{P}_{\text{assoc}}) \tag{3}$$

These associations effectively translate abstract properties into concrete, observable visual cues.

Filtering via Semantic Expert. To mitigate potential hallucinations within $\mathcal{A}$, we employ a frozen semantic expert Φ_{Sem} (e.g., Visualized BGE) to perform strict evidence filtering. The joint image-context and each candidate association are embedded into a shared latent space:

$$\mathbf{z}_I = \Phi_{\text{Sem}}(I, C), \quad \mathbf{z}_{a_i} = \Phi_{\text{Sem}}(a_i) \tag{4}$$

We compute the cosine similarity to measure the contextual alignment of each association:

$$s_i = \text{Sim}(\mathbf{z}_I, \mathbf{z}_{a_i}) = \frac{\mathbf{z}_I^\top \mathbf{z}_{a_i}}{\|\mathbf{z}_I\| \|\mathbf{z}_{a_i}\|} \tag{5}$$

The top-K associations with the highest similarity scores are retained as the verified evidence set $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$.

Grounding via Grounding Expert. For each retained evidence element $e_j \in \mathcal{E}$, a visual grounding expert Φ_{Ground} (e.g., Grounding DINO) localizes its corresponding region within the image space:

$$R_j = \Phi_{\text{Ground}}(I, e_j) \tag{6}$$

where R_j denotes the bounding box coordinates. This step anchors abstract symbolic cues onto concrete spatial regions, enabling localized reasoning.

Evidence-based Verification. Finally, the grounded evidence and their corresponding spatial regions are fed back into the MLLM to determine whether they collectively and sufficiently support the figurative ground G:

$$y = \text{MLLM}\left(I, Q, E, T, G, \{(e_j, R_j)\}_{j=1}^{K}\right) \tag{7}$$

By aggregating multiple grounded cues, AGES enables robust reasoning over abstract properties, leading to more robust relevance estimation under literal distractors.

3.5 Cross-Instance Evidence Comparison

Ultimately, AGES aggregates the individual verification results to execute the final prediction. It comprehensively compares all options based on the sufficiency and contextual strength of the grounded evidence supporting both tenor and ground. The candidate image that best satisfies these joint hierarchical conditions is selected as the final predicted answer.

4 Experimental Setup

4.1 Dataset & Metrics

Our experiments are conducted on the IRFL benchmark (Yosef et al., 2023). As introduced in Section 1, Multimodal Figurative Language Detection evaluates a vision-language model's capacity to identify the image that best conveys the figurative meaning of a textual expression. Given a figurative query and four candidate images, the model must select the single correct visual match. For idioms, the distractors include 1~2 partially literal images. Similes feature two literal distractors: one depicting the tenor without satisfying the ground (or showing its antonym), and another depicting the vehicle. Metaphors include 1~3 partially literal distractors. Performance is evaluated using *accuracy*.

4.2 Baselines

We compare our method with both traditional vision–language pre-trained models and open-source multimodal large language models in order to provide a comprehensive and fair evaluation. We adopt CLIP (Radford et al., 2021) as the representative traditional VL-PTM. Although it is not trained for figurative text-image matching, it provides a meaningful reference point for assessing the gap between literal matching and figurative modeling. For multimodal large language models, we include Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-32B-Instruct, Qwen3-VL-8B-Thinking (Bai et al., 2023), InternVL3-8B, InternVL3-14B (Chen et al., 2024), and GLM-4.1V-9B-Thinking (Team et al., 2026). Specifically, for the **Instruct** MLLMs, we adopt three settings:

- **Scoring.** We require the model to process each candidate image separately: analyzes the figurative language, evaluates the image–text alignment, and assigns a matching score; the highest-scoring image is selected as the answer.
- **Answer.** All candidate images are input together, and the model directly selects the image that best matches the figurative language.
- **CoT.** All images are also provided together, but the prompt additionally includes definitions of figurative language and explicit step-by-step reasoning instructions: analyze the figurative language, examine each image, and then compare all images to determine the final answer.

Table 1: Results on Multimodal Figurative Language Detection Task. Bold values indicate the best performance, while underlined values indicate the best performance of the same model under different settings.

	Models	**Idiom**		**Metaphor**	**Simile**	
		w/o Know	w/ Know		Closed	Open
-	CLIP	0.1600	0.2600	0.2342	0.4513	0.3791
Scoring	Qwen2.5-VL-7B-Instruct	0.3916	0.5300	0.4024	0.3430	0.2419
	InternVL3-8B	0.4500	0.5350	0.4835	0.4621	0.3863
	InternVL3-14B	0.5067	0.5350	0.5435	0.4982	0.4224
	Qwen2.5-VL-32B-Instruct	0.6033	0.6950	0.6637	0.5451	0.5018
Answer	Qwen2.5-VL-7B-Instruct	0.3533	0.4167	0.4464	0.5054	0.3177
	InternVL3-8B	0.4350	0.5017	0.4044	0.3622	0.2142
	InternVL3-14B	0.4600	0.5367	0.5085	0.4826	0.4055
	Qwen2.5-VL-32B-Instruct	0.6083	0.6783	0.7137	0.5181	0.4236
CoT	Qwen2.5-VL-7B-Instruct	0.4150	0.5350	0.4114	0.3448	0.2437
	InternVL3-8B	0.4650	0.5537	0.4515	0.4765	0.3213
	InternVL3-14B	0.4817	0.5817	0.5505	0.5451	0.4465
	Qwen2.5-VL-32B-Instruct	0.5833	**0.7350**	0.7377	0.6462	0.5211
Thinking	Qwen3-VL-8B-Thinking	**0.6150**	0.6850	0.6967	0.4296	0.4188
	GLM-4.1V-9B-Thinking	0.5850	0.6700	0.6667	0.6968	0.4729
AGES	Qwen2.5-VL-7B-Instruct	0.4250	0.6233	0.6504	0.7581	0.6715
	InternVL3-8B	0.5083	0.5550	0.6592	0.8015	0.6932
	Qwen2.5-VL-32B-Instruct	**0.6150**	0.7309	**0.8108**	**0.8484**	**0.7906**

For **Thinking** models, we similarly provide all candidate images with definitions of different types of figurative language, but do not impose an explicit CoT procedure, as we observed that explicitly constraining the reasoning process of Thinking models can adversely affect performance. More implementation details can be found in Appendix B.

5 Results

5.1 Main Results

The results for the Multimodal Figurative Language Detection task are presented in Table 1. Following the error analysis presented in Figure 1 (b), we evaluate the models on idiom detection under two settings: (1) without external knowledge integration (w/o Know), and (2) with external idiom knowledge incorporated from dictionaries (w/ Know). From the experimental results, we draw the following key observations:

First, CLIP benefits from the explicit presence of the tenor in similes for text–image matching, but shows a substantial performance gap on idiom and metaphor. It struggles with higher abstraction and non-compositional meanings, where MLLMs show clear advantages. Second, providing all images simultaneously performs better overall than scoring each candidate image separately. This may be because the best option is often determined relatively rather than absolutely. Meanwhile, scoring each image independently does not eliminate the effect of distractors, suggesting that literal bias is an inherent issue of the model. Third, despite being the most explicit figurative form, similes yield the worst performance for MLLMs under Answer and CoT settings. We observed that MLLMs often misidentify the semantic focus, treating the vehicle rather than the tenor as central.In addition, Thinking models tend to perform better on the Multimodal Figurative Language Detection task, where 8–9B Thinking models can reach or even exceed the performance of 32B Instruct models. Overall, the results demonstrate the effectiveness of our

Table 2: Ablation study on Multimodal Figurative Language Detection Task based on Qwen2.5-VL-7B-Instruct.

	Idiom		Metaphor	Simile	
	w/o Know	w/ Know		Closed	Open
Full AGES	0.4250	**0.6233**	**0.6504**	**0.7581**	**0.6715**
w/o evidence region	0.4150	0.6083	0.6486	0.7545	0.6715
w/o evidence filtering	0.4183	0.6075	0.6471	0.7509	0.6643
w/o association-guided evidence search	**0.4513**	0.5600	0.6291	0.7564	0.6559
w/o component-wise figurative verification	0.4167	0.5150	0.4114	0.3213	0.2347

proposed AGES, which yields significant performance gains on Multi-modal Figurative Language Detection.

5.2 Ablation Study

Table 2 present the results of the ablation study of our proposed AGES framework. All experiments are conducted using Qwen2.5-VL-7B-Instruct as the backbone model. The *w/o evidence region* variant removes the RoI predicted by the grounding expert and provides only textual evidence to the model. The *w/o evidence filtering* variant removes the evidence filtering and top-K (K=3) selection step, such that all candidate visual elements generated by the association module are directly fed into the model. The *w/o association-guided evidence search* variant completely disables the AGES module, requiring the model to perform figurative matching solely through tenor and ground verification without any explicit evidence search. The *w/o component-wise figurative verification* variant removes the explicit decomposition of figurative queries into tenor, vehicle, and ground, and omits the whole component-wise figurative verification process. Overall, the results demonstrate the effectiveness of the AGES framework. We summarize several key observations below.

Evidence regions facilitate reasoning. Providing the model with only textual evidence labels results in a noticeable performance decline. This suggests that incorporating visual evidence regions helps the model focus on figurative-related cues and supports more accurate reasoning.

Evidence filtering is essential. Removing the evidence filtering step leads to a clear performance drop, indicating that unfiltered evidence introduces substantial noise and increases the risk of hallucinated or weakly relevant visual cues.

The benefit of AGES depends on correct figurative interpretation. Without external idiom knowledge, the model often misinterprets idiomatic meanings, leading to evidence retrieval aligned with incorrect grounds and degraded performance. In contrast, when external idiom knowledge is incorporated, AGES is able to retrieve visually grounded evidence that supports the correct figurative interpretation, leading to consistent performance gains.

Component-wise figurative verification are crucial. By explicitly extracting the tenor and ground as verification conditions, AGES constrains candidate images to satisfy both figurative components simultaneously. This strategy is substantially more effective than direct image-text matching, which lacks such structured constraints.

5.3 Case Study

5.3.1 When AGES Works

Figure 3 presents an example in which applying AGES leads to improved model performance. Without AGES, the model concludes that "The soldier is not directly compared to a lion, nor is the soldier's appearance or behavior described in terms of lion-like qualities such as strength and dominance." After incorporating AGES, however, the model is able to verify that the armed soldier standing before a fiery explosion can serve as valid visual evidence of strength and dominance, which aligns well with the metaphorical meaning of *"a lion on the battlefield."* This example demonstrates AGES's ability to facilitate more grounded and semantically aligned reasoning when the metaphor can be correctly interpreted and supported by visual cues.

5.3.2 When AGES Fails

Figure 4 illustrates two common failure modes: (1) **Misinterpretation of figurative language.** The idiom "sick as a parrot" means "extremely disappointed" rather than a literal state of illness. When

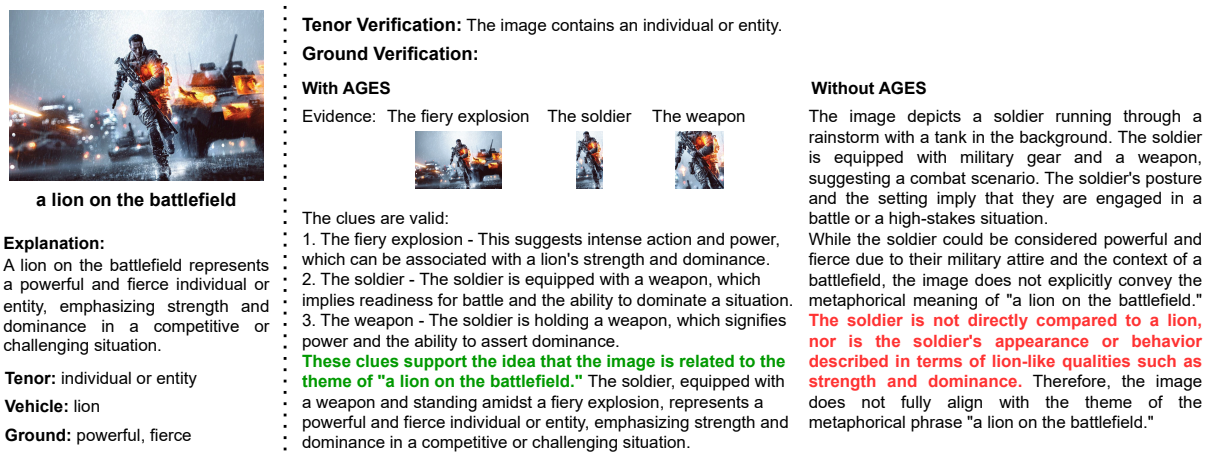
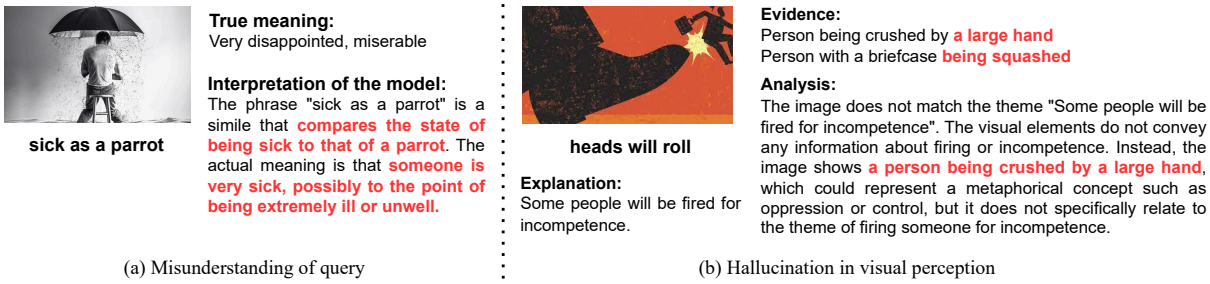


Figure 3: A case where AGES facilitates the model's matching of the image and the figurative query.



(a) Misunderstanding of query

(b) Hallucination in visual perception

Figure 4: Examples of two failure modes.

the model incorrectly interprets the meaning of the phrase, all subsequent reasoning is built upon a flawed understanding, ultimately resulting in an erroneous prediction. This case highlights that AGES critically relies on correct interpretation of figurative language; failures in figurative language understanding can significantly undermine its effectiveness.

(2) **Visual hallucination.** In Figure 4 (b), the candidate image depicts a person with a briefcase being "kicked away" by a large foot, symbolizing being fired. However, the model mistakenly identifies the foot as a hand and interprets the scene as a large hand attempting to crush a person, thereby failing to capture the intended metaphorical meaning. This example indicates that the effectiveness of AGES is also closely tied to the underlying visual perception capabilities of current MLLMs, and visual misrecognition can directly limit its performance.

6 Conclusion

In this work, we investigated the challenge of multimodal figurative language detection (formulated here as a multi-instance matching task), which requires models to transcend surface-level, literal alignments to capture abstract, non-literal meanings. Through systematic evaluation across diverse foundation models, we demonstrated that both VL-PTMs and MLLMs exhibit a pervasive literal bias when confronted with carefully designed literal distractors. Our analysis revealed that these failures stem not merely from insufficient world knowledge, but more fundamentally from dual bottlenecks: an inability to decode abstract visual symbolism and a failure to prioritize figurative relevance over literal overlap. To overcome these limitations, we introduced AGES, a training-free Association-Guided Evidence Search framework that explicitly decomposes figurative expressions and leverages an association mechanism to anchor model attention onto grounded visual evidence. Extensive experiments demonstrate that AGES consistently improves performance across various backbone models. Our findings underscore the boundary of current vision-language models in abstract reasoning, suggesting that structured, evidence-level association is a highly promising avenue for advancing multimodal understanding.

Limitations

While our proposed framework, AGES, demonstrates substantial improvements in suppressing literal bias, several inherent limitations remain, which we delineate below as scopes for future investigation.

Scope of Investigation. As discussed in Section 5.3, the primary objective of this work is to mitigate the pervasive *literal bias* in MLLMs–specifically, the phenomenon where a model correctly comprehends a figurative expression yet inadvertently favors images depicting purely literal components. Consequently, two orthoganol challenges fall outside the scope of this paper:

- **Figurative Language Interpretation**: Fully resolving highly complex or culturally-specific non-literal text remains a bottleneck. However, our experimental results indicates that integrating external knowledge via Retrieval-Augmented Generation (RAG) could substantially bolster the correct interpretation of such figurative queries.

- **Visual Hallucination**: Hallucination is a notorious, unresolved vulnerability in current multimodal foundation models. While AGES incorporates an objective perception stage to anchor high-level reasoning, fully eliminating foundational visual hallucinations is beyond the scope of this framework.

Semantic Loss in Evidence Grounding. The core mechanism of our evidence search relies on projecting abstract figurative concepts onto concrete visual cues. Due to the inherent abstraction of figurative language, we operationalize this by first utilizing an MLLM to distill abstract properties into objective textual descriptions, and subsequently employing external visuals expert for spatial localization.

During this hierarchical pipeline, semantic loss is inevitable. The efficacy of the grounding process is fundamentally bounded by the language-guided localization capability of the grounding expert. Developing a seamless, zero-loss translation mechanism that bridges abstract figurative semantics directly with concrete spatial regions represents a highly promising and valuable direction for future research.

References

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.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Giuseppe Gallipoli and Luca Cagliero. 2025. It is not a piece of cake for GPT: Explaining textual entailment recognition in the presence of figurative language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9656–9674, Abu Dhabi, UAE. Association for Computational Linguistics.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

George Lakoff. 1993. *The contemporary theory of metaphor*, page 202–251. Cambridge University Press.

George Lakoff and Mark Johnson. 2024. *Metaphors we live by*. University of Chicago press.

Yu Xi Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024. EmbodiedBERT: Cognitively informed metaphor detection incorporating sensorimotor information. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16868–16876, Miami, Florida, USA. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7069–7086.

Xingyuan Lu, Yuxi Liu, Dongyu Zhang, Zhiyao Wu, Jing Ren, and Feng Xia. 2025. Emometa: A multimodal dataset for fine-grained emotion classification in chinese metaphors. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3080–3083.

Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2025. Tug-of-war between idiom's figurative and literal meanings in llms. *ArXiv*, abs/2506.01723.

Inike Tesiana Putri and Widyastuti Purbani. 2018. Metaphors in the novel of and the mountains echoed by khaled hosseini using a stylistic analysis.

Wenhao Qian, Zhenzhen Hu, Zijie Song, and Jia Li. 2025. Concept drift guided layernorm tuning for efficient multimodal metaphor identification. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 1100–1108.

Le Qiu, Emmanuele Chersoni, and Aline Villavicencio. 2025. ChengyuSTS: An intrinsic perspective on Mandarin idiom representation. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 1–12, Suzhou, China. Association for Computational Linguistics.

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

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-flute: Visual figurative language understanding with textual explanations. *CoRR*.

Meghdut Sengupta, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, Debanjan Ghosh, and Henning Wachsmuth. 2025. Investigating the impact of conceptual metaphors on LLM-based NLI through shapley interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17393–17403, Suzhou, China. Association for Computational Linguistics.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Yuan Tian, Nan Xu, Wenji Mao, and Daniel Zeng. 2023. Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Yuan Tian, Ruike Zhang, Nan Xu, and Wenji Mao. 2024. Bridging word-pair and token-level metaphor detection with explainable domain mining. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13311–13325, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding

challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101.

Cheng Yang, Puli Chen, and Qingbao Huang. 2024. Can ChatGPT's performance be improved on verb metaphor detection tasks? bootstrapping and combining tacit knowledge. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1016–1027, Bangkok, Thailand. Association for Computational Linguistics.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Jenny Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. 2025. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. 2023. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6141–6154.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. Multimet: A multimodal dataset for metaphor understanding. 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 3214–3225.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1483–1497, Toronto, Canada. Association for Computational Linguistics.

A More Details about the Preliminary Experiments

Here are the definitions of the failure modes mentioned in our preliminary studies.

Phrase misinterpretation. This issue mainly occurs in idiom samples: when the model lacks idiom-related knowledge, it may misinterpret the phrase, and the subsequent analysis will continue based on that incorrect understanding.

Image misinterpretation. This issue can be divided into two cases. In the first case, visual hallucination causes the model to misinterpret the image. In the second case, no hallucination occurs: the model correctly understands the objective content of the image, but fails to infer the figurative meaning conveyed by it.

Preferring Literal Distractors. In this case, the model correctly interprets the objective content of the image and, to some extent, correctly explains how the image conveys the figurative meaning of the phrase. However, when making the final choice, the model incorrectly judges that an image containing only literal elements better matches the figurative meaning of the phrase.

Others. This category includes cases such as refusal to answer, incorrect output formatting that prevents the regular expression from extracting the answer, and situations where the final answer is inconsistent with the preceding analysis.

Some erroneous samples involve more than one type of error. In such cases, we categorize them according to the priority order: Phrase Misinterpretation, Image Misinterpretation, and Preferring Literal Distractors. Figure 5, Figure 6, Figure 7,

Figure 8 illustrate several examples of the error types discussed above.

B Implementation Details

All experiments are implemented in Python 3 using PyTorch 2. Models are are loaded via the Hugging Face Transformers library (Wolf et al., 2020), which provides unified interfaces for model loading, preprocessing, and generation. For Thinking models, we adopt vLLM (Kwon et al., 2023) to accelerate autoregressive decoding and improve inference efficiency. Experiments are conducted on NVIDIA A800 GPUs. Due to its larger parameter scale, Qwen2.5-VL-32B-Instruct is deployed across two GPUs, whereas all other models are executed on a single GPU. Due to hardware limitations, we did not evaluate AGES on OpenGVLab InternVL3-14B, as its dynamic high-resolution strategy produces a large number of image tokens and leads to out-of-memory (OOM) issues.

For our AGES framework, the Semantic Expert employs the visualized bge-m3 model, while the Grounding Expert utilizes grounding-dino-base. During evidence grounding, when multiple bounding boxes are detected, non-maximum suppression (NMS) is applied to remove redundant boxes with an Intersection-over-Union (IoU) threshold of 0.6. We did not

To ensure result stability, each experiment is conducted three times with different random seeds, and the final results are obtained by averaging over the three runs. Unless otherwise specified, all models use their default generation configurations, and no task-specific hyperparameters are tuned.

C Performance on Multimodal Figurative Retrieval

Yosef et al. (2023) introduced another task named Multimodal Figurative Retrieval, which requires the model to rank a larger image pool–containing both figurative and partially literal images–based on a figurative query. The objective is to rank the images such that those expressing the figurative meaning are ranked higher than the partially literal ones. The evaluation metric is *precision@k*, where k denotes the number of figurative images in the candidate set.

We further evaluate ARGS on Multimodal Figurative Retrieval with a small adaption: After aggregating the verification evidences, each image is assigned a scalar relevance score $S \in [1, 10]$. Then all candidate images are ranked by S. Table 3 present the results on the Multimodal Figurative Retrieval task. Overall, AGES achieves superior performance. However, when Qwen2.5-VL-32B is used as the backbone model, its performance on idiom retrieval is relatively limited. It remains unclear whether this limitation stems from the intrinsic capabilities of the backbone model itself, and this issue may be further investigated in future work.

sick as a parrot

Phrase explanation:
Very disappointed; miserable

Response:
Step 1. Phrase Analysis
The phrase "sick as a parrot" is a simile. It means someone is very sick, often used humorously to indicate extreme illness. The comparison emphasizes the severity of being sick by likening it to the state of a parrot, which can be ill or unwell.

Step 2. Image Analysis
- **Image A**: A person sitting under an umbrella in the rain. This image does not depict a parrot or illness; it shows a person in a rainy environment.
- **Image B**: A colorful parrot being fed from a bottle. This image directly depicts a parrot, which aligns with the phrase's metaphorical meaning.
- **Image C**: A group of people in traditional attire, possibly in a historical or cultural setting. This image does not relate to the phrase at all.
- **Image D**: An interior view of a restaurant with tables and chairs. This image is unrelated to the phrase and does not depict a parrot or illness.

Step 3. Final Decision
Answer: B

Figure 5: A bad case of caused by phrase misinterpretation. The model incorrectly interpreted “sick as a parrot” as “very sick”, when in fact it is an idiom meaning “extremely disappointed”.

in your face

Phrase explanation:
An exclamation of derision and contempt

Response:
Step 1. Phrase Analysis
The phrase "in your face" is an idiom that means to confront someone directly, often in a rude or aggressive manner. It implies a direct and confrontational action where the person is literally or figuratively in front of you.

Step 2. Image Analysis
- **Option A:** The image shows a light bulb with a smiling face pointing upwards. This image does not depict confrontation or rudeness; instead, it conveys a positive and friendly message.

- **Option B:** The image shows a person with their hand raised, as if to block or push something away. This could symbolize resistance or refusal, but it doesn't specifically convey confrontation or rudeness in the context of being "in your face."

- **Option C:** The image shows a toy horse with a painted face. There is no indication of confrontation or rudeness here; it is simply a decorative item.

- **Option D:** The image shows a close-up of a person's face with their hand raised, as if to block or push something away. This image aligns well with the idiom "in your face," as it suggests a confrontational gesture.

Step 3. Final Decision
Answer: D

Figure 6: A bad case caused by image misinterpretation. The model hallucinated significantly in its description of option D.

take to the hills

Phrase explanation:
To flee or vanish; to run away

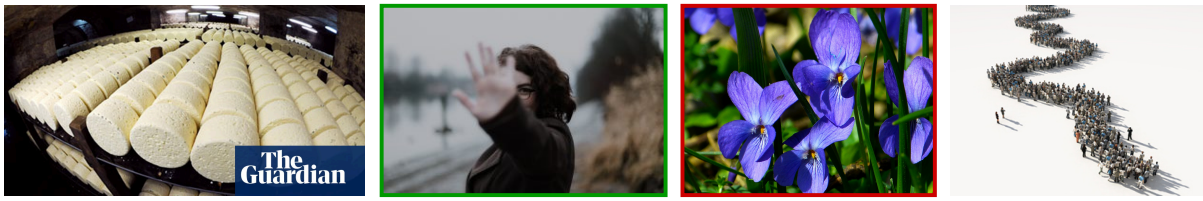
Response:
Step 1. Phrase Analysis
The phrase "take to the hills" is an idiomatic expression. It means to leave a place quickly, often in a hurry or to escape danger. The metaphorical meaning is to flee or retreat to a safer or more remote location.

Step 2. Image Analysis
A. The image shows a snowy forest scene with trees covered in snow and sunlight shining through the branches. This image depicts a natural, serene environment but does not convey the idea of fleeing or escaping danger.
B. The image shows a person standing by a car, possibly waiting for someone or something. There is no indication of fleeing or escaping danger in this image.
C. The image shows a person holding a knife, which could symbolize danger or aggression but does not directly relate to the concept of fleeing or escaping.
D. The image shows a person floating in space with various icons around them, which could represent a sense of being disconnected or removed from reality. However, it does not clearly depict the act of fleeing or escaping danger.

Step 3. Final Decision
Answer: A.

Figure 7: A bad case caused by image misinterpretation. The model failed to correctly infer from the context that option B likely depicts someone hurriedly leaving with luggage, which matches the theme of “take to the hills”.

shrinking violet



Phrase explanation：
A very shy or timid person, who avoids contact with others if possible

Response:
Step 1. Phrase Analysis
The phrase "shrinking violet" is a metaphor. It refers to someone who is shy, timid, or retiring, often described as being like a violet flower that appears small and unassuming when compared to other flowers.

Step 2. Image Analysis
- **Option A**: This image shows rows of cheese wheels in a storage area. There is no connection to a violet flower or any metaphorical meaning related to shyness or timidity.

- **Option B**: This image depicts a person standing outdoors with their hand raised, possibly signaling "stop." The person appears to be alone and could symbolize isolation or shyness, but there is no direct reference to a violet flower or shrinking.

- **Option C**: This image shows a close-up of purple violets. The violets are small and delicate, which can metaphorically represent a shy or timid person. This image directly aligns with the metaphorical meaning of "shrinking violet."

- **Option D**: This image shows a large group of people forming a line. This image does not relate to the metaphorical meaning of a shy or timid person.

Step 3. Final Decision
Answer: [C]

Figure 8: A bad case caused by preferring literal distractors. The model correctly identified that option B reflects the character's isolation or shyness, but it concluded that it does not fit the theme of "shrinking violet" because it does not reflect flowers. Instead, it mistakenly considered a photo of violet flowers (a literal image) to match the figurative meaning of "shrinking violet".

Table 3: Results on Multimodal Figurative Retrieval task. Bold values indicate the best performance, while underlined values indicate the best performance of the same model under different settings.

	Models	**Idiom**		**Metaphor**	**Simile**	
		w/o Know	w/ Know		Closed	Open
-	CLIP	0.4376	0.4376	0.2254	0.3944	0.3062
Answer	Qwen2.5-VL-7B-Instruct	0.5655	0.6124	0.3831	0.2209	0.1276
	InternVL3-8B	0.4550	0.5100	0.3874	0.3718	0.2274
	InternVL3-14B	0.6016	0.5999	0.4764	0.2944	0.2717
	Qwen2.5-VL-32B-Instruct	0.6949	0.7316	0.6280	0.3439	0.3218
CoT	Qwen2.5-VL-7B-Instruct	0.6008	0.6722	0.4752	0.2790	0.2137
	InternVL3-8B	0.5783	0.6441	0.4749	0.3036	0.2203
	InternVL3-14B	0.6299	0.6356	0.4739	0.2967	0.2699
	Qwen2.5-VL-32B-Instruct	**0.7307**	**0.7654**	0.6267	0.4261	0.3699
Thinking	Qwen3-VL-8B-Thinking	0.6297	0.6570	0.5062	0.2051	0.1543
	GLM-4.1V-9B-Thinking	0.6264	0.6916	0.4464	0.3082	0.2008
AGES	Qwen2.5-VL-7B-Instruct	0.6232	0.6769	0.6417	0.7454	0.6193
	InternVL3-8B	0.6789	0.7126	0.6202	**0.7750**	0.6641
	Qwen2.5-VL-32B-Instruct	0.6862	0.7175	**0.6741**	0.7534	**0.6665**