On Vision Language Models' Retrievals Based on Figurative Expressions

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) hold much promise for image understanding by merging text and images, possibly through training on web-scale data. Still, there are issues with the alignment of the text and images. This study explores the effect of linguistic variety on image retrieval in modern VLMs. The study publishes experimental results on the effects of linguistic varieties, highlighting possible areas for improvement. We utilize established techniques for measuring retrieval and text-image alignment. We introduce a use case where VLMs could be presented with a figurative expression to find a matching image with day-to-day objects that it has knowledge of. We found the problem challenging for VLMs. Although we were unable to provide a solution, we highlight a technique by using the properties of the concepts.

1 Introduction

Over the last decade, multi-modal Vision Language Models have been increasingly deployed, demonstrating their practical utility. Although the models have improved in a few short years, certain nuances in language are still being missed. In fact, (Yuksekgonul et al., 2022) observes that these models do not provide evidence of understanding nuances in language. For example, *the horse is eating the grass* and *the grass is eating the horse* seem to make no difference to these models. Typically, these models could be tested for understanding a query by observing their retrieval accuracy. The retrieval of an image based on the two expressions should be vastly different; however, apallingly, they are not – often resulting in fetching the same image.

(Parcalabescu et al., 2022) examines these lapses in more detail by probing for some common sense knowledge like the existence of certain objects in an image and the number of objects present in the image, in addition to checking plurality, coreference, and spatial relation. Of late, the VLMs are being used more abundantly with more focus on increasing their performance, which is why we perform the tests on the latest models, hoping to gain some insights.

In another vein, these models seem to perform well enough on the common and simple usage of the language. Query the model on *a black cat standing in front of a motorcycle*, the model will likely fetch relevant images (Fig. 1) because these models have been exposed to everyday objects in training. However, a reformulated query using metaphor *A small shadow standing guard before a sleeping thunderbolt.*, leads the model to retrieve some random images.

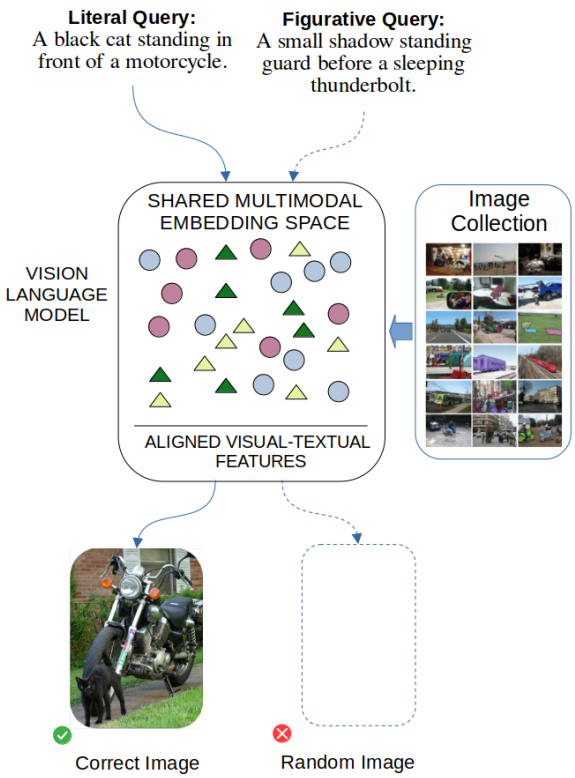

Figure 1: Metaphor as query to VLM to retrieve matching image.

There has been some interest in interpreting fig-

urative language in textual language models, also. In fact, (Jing et al., 2025) explores the representation of metaphors, synecdoche, personification, and other linguistic phenomena in language models. We extend the work in the visual language model domain. In addition to re-testing the existing benchmark mentioned above (Parcalabescu et al., 2022), we test the models for the following linguistic constructs: epistemic, euphemism, expressive, metaphor, personification, politeness, synecdoche, and topic comments.

Although much work has been done on the detection of such, no work has been done that does the reverse mapping – in other words, from metaphor to literal. The utility of such mapping could be useful to a user to find a suitable image for a metaphoric expression, to say the least.

The VLMs we subject to experiment come ready after being trained jointly on text and image data. For probing them, we employ common scores used to gauge perplexity or vector similarity, depending on the model's architecture. We find proof of ignorance about figurative expressions such as metaphors, synecdoche, and personification. We explore ways to mitigate the issues. Since working on multiple types of expressions is very laborious, we work only on metaphors and explore ways to improve retrieval efficiency. We utilize theories such as Conceptual Metaphor Theory(CMT) (Lakoff, 1993) and well-known datasets such as MS-COCO (Chen et al., 2015) to demonstrate the ideas.

In short, our contributions are –

- Probe the models for figurative expressions
- Draw insights by observing proof of activation in the models
- Suggest ways to minimize metaphor to literal discrepancy
- Test an existing benchmark on newer models

2 Background

2.1 Vision Language Models

Modern VLMs have been trained jointly on images and corresponding text. A base architecture that all models follow is ViT (Dosovitskiy et al., 2020). For training ViT, images are broken into patches and fed into a transformer model to predict a single word descriptive of the image. This imparts knowledge of concepts in the image patches. Datasets that were used are ImageNet (Deng et al., 2009) and JFT-300M (Google's internal) (Sun et al., 2017). Building on the success of ViT, the CLIP model (Radford et al., 2021) has been built. This is a popular model for predicting natural language captions instead of just a single word. Images are jointly trained on full captions using cosine similarity in the network's loss calculation. The model has been trained on web-scale data, which makes it attractive to a researcher.

While the earlier models perform just retrieval, later models and current state-of-the-art models have enhanced capabilities. SigLip and SigLip-2 (Tschannen et al., 2025) have been introduced to incorporate multi-lingual capabilities. LLaVA (Liu et al., 2023) and its subsequent versions have language generation abilities. These models have been trained with generative language models such as Llama (Dubey et al., 2024), which enables them to generate description stories given an image. QWEN (Yang et al., 2024), Gemma (Team et al., 2024) are also examples of such generative models.

One particular weakness of the models is that they often lack grounding, as modern generative models tend to be either vision or textual. For example, counting the number of objects (Paiss et al., 2023), determining the relative ordering of objects (what object is obscured by another) in an image (Qi et al., 2025), or finding how objects are spatially located relative to each other (Thrush et al., 2022), leads to confusion in the models.

Some modern models, such as KOSMOS-2 (Peng et al., 2024), GroundingDino (Liu et al., 2024), attempt to overcome these by using grounding information in the training. To our knowledge, none of the models have been trained to handle the advanced linguistic features that this paper focuses on.

2.2 Linguistic Features in VLMs

Some works point out subtle weaknesses in these models by making minimal changes in the text and observing the output change. (Parcalabescu et al., 2022), for example, does actant swap. Given an image where a man is seen shouting at a woman, if the model is given a misleading caption indicating the woman is shouting at the man, the model's belief does not register the swap. For brevity, we are skipping all the types of linguistic foil they have made. We, however, use their foils in our experiments on modern VLMs. A key takeaway is that the models seem to act as though the training captions are a bag of words without much deep

semantic meaning or world knowledge.

While the (Parcalabescu et al., 2022) limits itself to basic textual manipulations, of late, research focuses on deep semantic meaning and linguistic constructs such as using figurative language. (Jing et al., 2025) performs interpretability on language models (not VLMs) on syntactic constructs, pragmatic functions, rhetorical expressions such as politeness, epistemic, euphemism, personification, synecdoche, and metaphors. This leads us to check VLM's capabilities along these dimensions.

2.3 Metaphor Detection

Metaphors have been widely studied both in linguistic and natural language processing arenas. A key technique is Metaphor Identification Procedures (Group, 2007). Based on the technique, the NLP community has devised ways to identify metaphors using contextual clues and identifying out-of-place word usage. Several seminal NLP works, such as FrameNet (Baker et al., 1998) and VerbNet (Shi and Mihalcea, 2005), help identify such semantic shifts since these knowledge bases carry much information about regular usages of words. A contextual embedding-based work, FrameBERT(Li et al., 2023), has also been developed to identify metaphors.

Researchers can use several tagged corpora to check their metaphor detection methods' accuracy and train their models. Some popular ones are MetaNet(Dodge et al., 2015), LCC-corpus (Mohler et al., 2016), MIPVU (Krennmayr and Steen, 2017). In these datasets, a metaphoric sentence is given, and an indication of what words are metaphoric is pointed out. Some even have a number signaling the extent of the metaphor. Our study suggests that most of these resources are focused on one special domain, such as news articles, issues related to governance, etc.

Conceptual Metaphor Theory(CMT) argues that metaphors could be thought of as using a source concept to describe a target concept (Lakoff, 1993). In "A career is a journey", *journey* is the source domain and *career* the target. The source is used to impute some properties to the target – *journey* is a space that can be traveled by *career*. CMT proposes that a metaphor can be understood as mapping from the source to the target domain.

Another study (Bowdle and Gentner, 2005) points to the similarity between target and source words in metaphors. According to this study, similarity helps in metaphor interpretation. We find ConceptNet (Speer et al., 2016), as a popular ontology of concepts, interesting for interpreting a metaphor since properties for concepts could indicate similarity. In addition, ConceptNet contains concepts and properties in a connected graph structure. Unlike other metaphor resources, ConceptNet contains references to day-to-day concepts as well. This is crucial because our work focuses on retrieving day-to-day objects from a metaphorical expression.

2.4 Embeddings

We experiment with several word embeddings – Word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), and Numberbatch (Speer et al., 2016). Word2vec and Glove are considered static embeddings as they primarily represent words in the vector space without distinguishing between multiple senses. The former two have been constructed based on the distribution of words. Numberbatch, on the other hand, has some relational information based on the knowledge graph of ConceptNet.

3 Methods

3.1 Probing Model Beliefs

Measuring beliefs of VLMs is what we focus on. The black-box model of neural networks is commonly probed for their beliefs by observing their activations and their outputs. One way to measure the belief of the black-box VLMs is to calculate retrieval efficiency. The accepted way to measure retrieval accuracy is Mean Reciprocal Rank (MRR) (Voorhees, 1999).

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{1}$$

, where $|Q|$ is the the number of queries and rank_i is the position of the relevant item.

We use cosine similarity between the vector representation of a query text and retrieved images as a measurement for computing MRR.

In addition, VLMs could be probed for their internal beliefs on a text and image pair based on cosine similarities. In some models, such as CLIP, we can determine the vector representations to apply cosine similarity. However, for generative models, such as LLaVA, no such vectors could be found that represent both entities. These generative models repeatedly generate tokens updating belief and

vector representaion as each single step, making it difficult to capture overall belief in a single vector. The recommended way is to use Negative Log Likelihood (NLL) (Goodfellow et al., 2016) to measure perplexity, which we utilize to find the model's belief about alignment between a text and image pair, Fig. 2.

$$\mathrm{NLL}(\theta) = -\sum_{i=1}^{n} \log P(x_i \mid \theta) \qquad (2)$$

, where x_i stands for the token generated at step i and θ stands for model parameters.

These performance measures have no absolute values. We rather use their relative values to other samples that we present to the models.

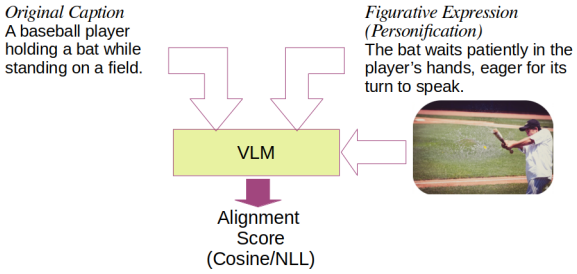

Figure 2: Alignment scoring of figurative vs cosine similarity or NLL.

3.2 Matching Metaphor Text with Literals

Based on our initial findings that show models lack in aligning figurative expressions with images, we attempt to find ways to mitigate this. As a test case, we consider metaphors. Our broad goal is to find what it would take to find an image based on a metaphoric sentence rather than the literal text that the model was originally trained on.

We reduce our problem to a more manageable lexical unit matching task. We extract the main nouns that are used to express the metaphor as well as the corresponding literal word. For example, in the following, *tower* and *refrigerator* are the literal and metaphor, respectively –

Literal Expression: a cat sitting on top of a blue **refrigerator**
Metaphoric Expression: A gargoyle keeping watch from a frozen azure **tower**

We randomly collect 239 captions from the MS-COCO dataset. We further create figurative expressions using an LLM based on the captions. We choose a literal word and a metaphorical word from the caption and figurative expression, respectively. Thus, each pair of our experiment contains a literal and a metaphorical lexical item with part of speech Noun. We find that the number of unique metaphor words is 180 because some words have been used in multiple metaphor formulations. We also collect 2,235 random nouns to use as test words. For each metaphor in the pair, we subject a model to locate the corresponding literal word from the 2,000+ words to test whether it understands the mapping between metaphor and literal lexical unit. The use of metaphors as a substitute for literal words or, as per the CMT metaphor theory and (Bowdle and Gentner, 2005), using source to express target, suggests that the source (metaphor) must share some common properties with the target (literal). We can think of the above example as a reformulated metaphoric expression "The refrigerator is a tower," even though we have two sentences instead of one. Nonetheless, the source *tower* is being used to impart a "tall" property to the target *refrigerator*. We use ConceptNet to locate properties common to both. We surmise that, being a rich source of information, ConceptNet will let us locate common properties. However, the ontology is found to have too few immediate properties for some concepts to compare. However, the nodes are connected via "IsA" relationships to containers, which could be traveled any number of hops. This lets us utilize superordinate properties of the lexical units.

The other issue is that properties are not comparable using surface forms alone. The source *thunderbolt* has a property "loud". The corresponding target *motorcycle* has properties "produces noise", "noisy". Semantically, we could match them only by using vector representations of words. Therefore, a vector model such as Numberbatch that is built with properties in mind could be appropriate. We use it to find cosine similarities between metaphors and the 2,000+ words. We also compare the performance of vectors from Word2Vec and Glove models. We observe the rank of the matched literal word and calculate MRR for each model we try (Fig. 3).

In our initial experiment, we found that Numberbatch provides the best results. We extend our experiment by incorporating more property information from ConceptNet by varying hops. ConceptNet has relationships in the form <StartWord, Relation, EndWord>. In any *n*-hop, we locate all the relations associated with the metaphor, create

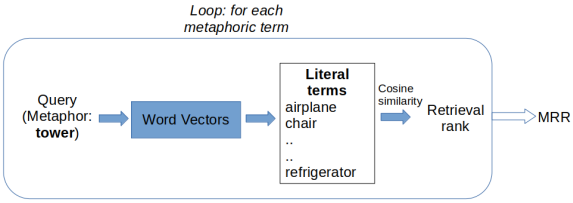

Figure 3: Alignment scoring of figurative vs cosine similarity or NLL.

a mean vector from the EndWords in all relations in that hop. The mean vector is used as the representation of the metaphor. This vector is used to retrieve the literals from the 2,000+ words.

3.3 Re-testing Benchmark

(Parcalabescu et al., 2022) tests models four years ago. Since then, VLMs have come a long way. We target some popular modern models and carry out the tests described in the research. Just as the authors suggested, we collect the acc_r score, which gives accuracy vs ground truth.

4 Experimentation and Results

4.1 Re-testing Benchmark

We carry out the testing and report the results on the latest VLMs like GEMMA-3, QWEN-3.5, LlAVA-1.6 and Llama-3.2 in addition to the original CLIP and Siglip-2 models (TABLE 1).

Table 1: Model Performance Comparison (Acc_R)

Model	Action	Coreference	Counting	Noun Phrase	Plurals	Relation
CLIP	0.72	0.51	0.61	0.89	0.56	0.64
SigLip	0.57	0.50	0.50	0.81	0.50	0.57
LLaVA	0.81	0.77	0.61	0.94	0.54	0.82
QWEN	0.83	0.84	0.69	0.96	0.58	0.90
GEMMA	0.58	0.51	0.60	0.86	0.51	0.81
Llama	0.86	0.75	0.70	0.97	0.59	0.91

4.2 Probing Model Beliefs for Linguistic Varieties

We formulate our probe by presenting a pair of sentences (Table 2). The first sentence describes regular day-to-day objects or events in plain English (literal expression), and the second sentence is a figurative expression. We measure the model's belief. If the model shows closer alignment between the first sentence and the associated image, we label the result as correct. Gauging the belief varies between models that are just retrieval-focused vs those that provide generative description. The former model's belief is measured by the vector dot product. For the rest of the models, we use NLL.

We use the MS-COCO dataset since the captions and images are publicly available. TABLE 3 and TABLE 4 show the results.

Table 2: Examples of presenting two alternative captions to the model

	Data
	Metaphor
Image:	[Image]
Literal Sentence:	The people are posing for a group photo.
Figurative Expression:	A sea of faces smiles for the camera.
	Synecdoche
Image:	[Image]
Literal Sentence:	A long train traveling through a train yard.
Figurative Expression:	Iron wheels roll steadily through the train yard.
	Personification
Image:	[Image]
Literal Sentence:	A green netted bed in a light-filled bedroom.
Figurative Expression:	The street fills itself with people, as though it has decided to host a gathering.

Table 3: Model Performance Comparison on Linguistic Categories (Accuracy)

Model	Epistemic	Euphemism	Existence	Expressive
CLIP	0.47	0.51	0.68	0.60
SigLip	0.42	0.41	0.57	0.50
LLaVA	0.19	0.42	0.51	0.69
QWEN	0.42	0.81	0.70	0.89
GEMMA	0.08	0.73	0.66	0.81
Llama	0.32	0.73	0.78	0.70

Table 4: Model Performance on Figurative Language and Pragmatics (Accuracy)

Model	Metaphor	Personification	Politeness	Synecdoche	Topic Comment
CLIP	0.87	0.63	0.38	0.94	0.57
SigLip	0.72	0.46	0.40	0.81	0.33
LLaVA	0.94	0.71	0.17	0.96	0.85
QWEN	1.00	0.91	0.31	0.99	0.90
GEMMA	0.94	0.61	0.27	0.88	0.68
Llama	0.95	0.74	0.26	0.99	0.82

Probe Result Analysis: Since we are exploring alignment between two captions – literal and figurative – against the ground truth image, the numbers in TABLE 4 indicate a high value for literal expression alignment. In other words, the more the model can discriminate between literal and figurative, the higher the value. We notice that figurative expressions, such as synecdoche and metaphor, have very high values. This is because the models have little notion of figurative language and always give a higher value for the literal counterpart. The farther figurative expressions are from the image's representation, the less its alignment value. We confirm this by analyzing individual cases in a scatterplot as below (Fig. 4) and find that the scores for figurative expressions are unusually low. This prompts us to investigate how we can bridge the gap so that

figurative expressions can retrieve the images that literal expressions can.

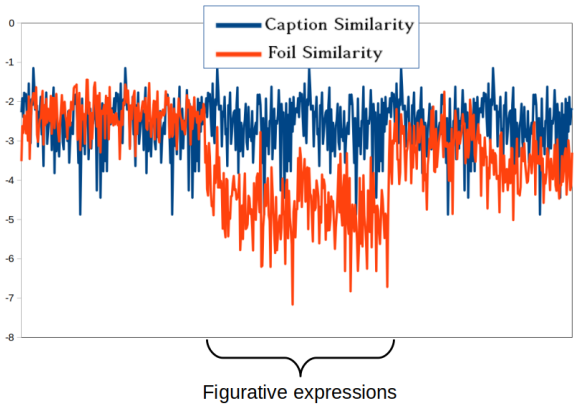

Figure 4: Plot of caption and foil similarities. Captions are literal expressions, and foils are figurative expressions. By similarity, we mean either cosine similarity or NLL.

4.3 Model Details

Here are the details of the model used in the experiment (TABLE 5):

Table 5: Comparison of Vision and Language Model Parameters Used in the Experiments. For 3, 4, 5, and 6, we use unsloth (Han and Contributors, 2024).

Sl	Model	Version	Parameters
1	CLIP	vit-base	≈ 151 million
2	SigLip	2	≈ 375 million
3	LLaVA	1.6	7 billion
4	QWEN	2.5-VL-3B-Instruct	3 billion
5	GEMMA	3	4 billion
6	Llama	Llama-3.2	11 billion

4.4 Addressing Shortcomings for Metaphoric Expressions

We break down both literal and metaphoric expressions and identify, by human supervision, the constituent source and target words as per the CMT theory. As the following image (Fig. 5) demonstrates, we collect target *tie* and source *formality* from the expressions. We collect 239 such pairs. For each pair, we find cosine similarities in the word vector models (TABLE 6).

We find that these vectors do not have a noticeable notion of similarity between literal and metaphor, understandably, because the vectors were trained on samples that do not place the words in similar contexts. The CMT theory suggests we use the source's properties to impart meaning to

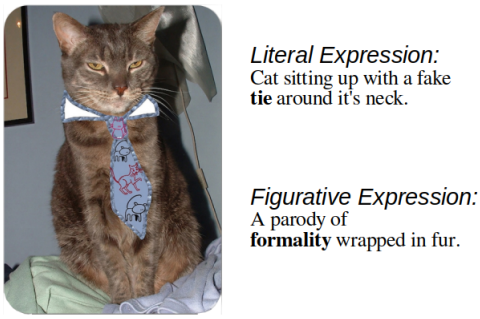

Figure 5: Lexical units that conveys target and source in metaphors – *tie* is the literal and *formality* is the metaphor.

Model	Random Pair	Metaphor Literal
Numberbatch	0.039	0.153
GloVe	0.154	0.195
Word2Vec	0.107	0.179

Table 6: Model Performance: Mean Cosine Similarity across different embedding models.

the target. For example, in "*The classroom was a **zoo**.*", the *classroom* is the target and *zoo* is the source to convey disorderliness. However, we just do not use a random word as a source. A close inspection reveals that we use *zoo* metaphorically, as opposed to a random word, because both *classroom* and *zoo* share some common properties, such as *contains_living_things* and *provides_protection*. In a similar vein, *motorcycle* and *thunderbolt* share common properties like *noisy* and *powerful*. We find that Numberbatch in TABLE 6 shows higher similarities. We further attempt to impute more property-based knowledge in our work by using ConceptNet and its property knowledge.

Experimentation to Retrieve Images for Metaphors: Building on the notion that two lexical units must share properties to be used in a literal-metaphor pair, we attempt to collect properties commonly associated with the items. ConceptNet has a substantial number of human-annotated properties. However, many concepts do not have many immediate properties. This would render matching futile. Therefore, we enhance the property information for the lexical units (Concepts) by collecting distant properties in ConceptNet. Fig. 6 shows MRR values at different hops. Fig. 6 reveals that Numberbach has the highest, although not a large, MRR value in 1-hop. This is because the creator of Numberbatch performed arithmetic vector alignment similar

to (Faruqui et al., 2015) to bring Word2Vec and ConceptNet's relations closer.

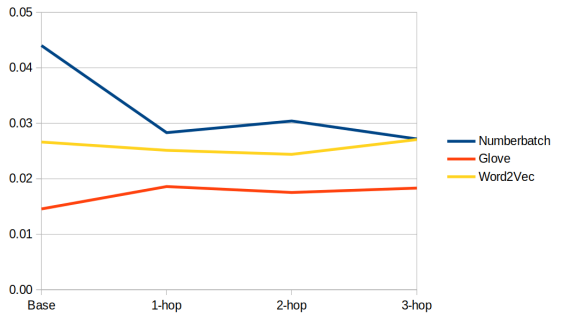

Figure 6: MRR for models at different hops. Base is the vector of the word as it is. The others are vectors for properties at n-hop away.

ConceptNet has some properties like "IsA", "DerivedFrom", and "Synonym" indicating the provenance of a concept. We use these as a signal of parenthood and use the parental properties for the children. In addition, a concept could be thought of as a collection of properties. To formulate a number representation of a lexical unit, we could use the vector mean of properties. As for initializing the vector, we could choose from many models. Fig. 7 shows how the mean cosine similarity score varies across hops for various models.

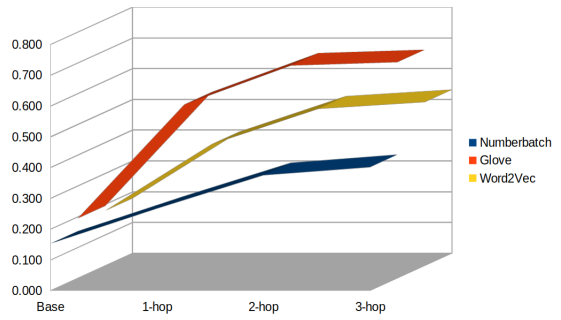

Figure 7: Average cosine similarities across hops between literal and metaphor words.

As a measure of accuracy, we resort to the popular retrieval ranking method, MRR. Given a metaphor's representation, we retrieve the most similar literal words present in our vector collection. We create more than 2,000 literal words to test our accuracy. TABLE 7 and 8 show the results.

Table 7: MRR Using Run-of-the-Mill Word Embeddings

Model	MRR
Numberbatch	0.0440
Glove	0.0146
Word2Vec	0.0266

Table 8: MRR by Varying Number of Hops in ConceptNet

Hops	Numberbatch	Glove	Word2Vec
One-hop	0.028	0.019	0.025
Two-hops	0.030	0.018	0.024
Three-hops	0.027	0.018	0.027

4.5 Discussions

We target a subset of metaphoric expressions, leaving a vast majority out of our experiment. Still, the task looks very challenging, some of which are as follows:

4.5.1 Combinatorial Explosion:

Finding common properties between a target and a source requires having access to property information. We took ConceptNet as the resource for property information. Below is a table (TABLE 9) that shows the properties in ConceptNet directly attached to the target and source at one hop.

Table 9: One-hop relations of motorcycle and thunderbolt

Concept: motorcycle		Concept: thunderbolt	
Relation	**EndConcept**	**Relation**	**EndConcept**
IsA	two	HasContext	heraldiccharge
	_wheeled		
	_motorized		
	_vehicle		
RelatedTo	motor	HasContext	paleontology
RelatedTo	scooter	HasContext	soccer
RelatedTo	two	RelatedTo	belemnite
RelatedTo	bike	RelatedTo	charge
		RelatedTo	daring
		RelatedTo	event
		RelatedTo	hero
		RelatedTo	horrific
		RelatedTo	irresistible
		RelatedTo	lightning
		RelatedTo	powerful
		RelatedTo	terrible
		RelatedTo	thunder
		RelatedTo	thunderclap
		RelatedTo	thunderstone
		RelatedTo	unexpected

Having no common properties between them, one is tempted to extend the exploration space by including 2 or 3 hop away nodes. In fact, when we do so we do arrive at some common properties (Fig. 8):

We find that finding common properties requires exploring distant properties. This raises a combinatorial explosion issue. For *thunderbolt*, the number of properties are given below (TABLE 10):

Table 10: Number of Properties Across Hops

1-hop	2-hop	3-hop
5	429	8,651

We also observe some key statistics in Concept-

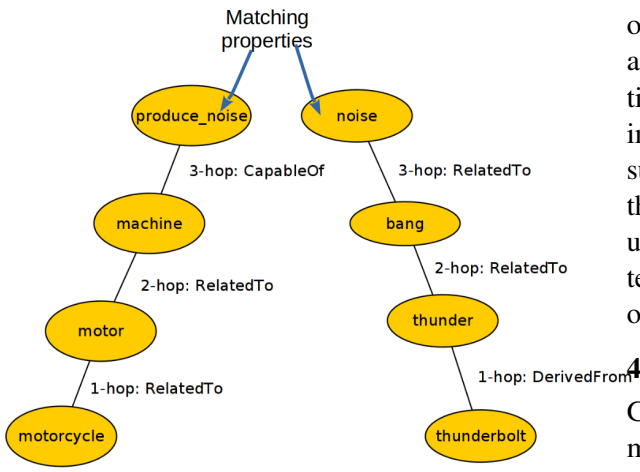

Figure 8: Two subgraphs for two concepts: *motorcycle* and *thunderbolt*. The third-hops show a common concept *noise*.

Net related to graph-theoretic concepts:

Table 11: This table is fixed in place

Number of edges (properties): 3,423,004
Number of nodes (concepts): 1,410,297
Average branching factor: 11.0
Approximate growth of the search space: $O(11^k)$

One observation is that the relations (TABLE 12) contained in ConceptNet do not fully conform to human expectations. For the "*Classroom was a zoo*", metaphor, it would have been ideal to have relations like "contains_living_things" and "provides_protection", which are not available in ConceptNet.

Table 12: Relations in ConceptNet

Relation	Relation	Relation
Antonym	HasContext	NotHasProperty
AtLocation	HasFirstSubevent	PartOf
CapableOf	HasLastSubevent	ReceivesAction
Causes	HasPrerequisite	RelatedTo
CausesDesire	HasProperty	SimilarTo
CreatedBy	HasSubevent	SymbolOf
DefinedAs	InstanceOf	Synonym
DerivedFrom	IsA	UsedFor
Desires	LocatedNear	dbpedia/capital
DistinctFrom	MadeOf	dbpedia/field
Entails	MannerOf	dbpedia/genre
EtymologicallyDerivedFrom	MotivatedByGoal	dbpedia/genus
EtymologicallyRelatedTo	NotCapableOf	dbpedia/influencedBy
FormOf	NotDesires	dbpedia/knownFor
HasA		dbpedia/language
dbpedia/leader	dbpedia/occupation	dbpedia/product

4.5.2 Surface-level Matching

As the Fig. 8 shows, the end concepts are not the same in form on the surface – produce_noise and noise. Therefore, semantic distance instead of surface-level matching must be considered. In addition, it is tempting to ignore "antonym" relations. We did that, but did not find considerable improvement. Even in modern word embeddings such as those we find in BERT (Devlin et al., 2018), the negation information as a concept is not well-understood (Ettinger, 2020). Therefore, it is a matter of discretion to exclude the antonym properties or to treat them just like any other property.

4.5.3 Future Direction

Given the cost and time needed to train the AI models, VLMs included, from the ground up, it is not practical to create training datasets containing figurative expressions and retrain. In addition, overuse will cause these expressions to lose their novelty and become too conventionalized to be considered novel (Bowdle and Gentner, 2005). Instead of retraining, we propose to work with the existing weights, mitigating the shortfalls as a processing step, something in the direction of research like (Wazni et al., 2024) attempt through vector arithmetic.

5 Limitation

The work focuses on English only.

6 Conclusion

We set out to understand how figurative expressions could be used to retrieve images. Along the way, we re-tested some existing benchmarks, introduced our own figurative expression test set, and attempted to solve a reduced problem through property matching. We find that the metaphor theories are sometimes contested, and there are many different ones. Furthermore, the existing metaphor datasets are inadequate for the day-to-day object retrieval that extant models specialize in. Although we fell short in providing a solution, we followed established measurement and methods, hoping to shed light on the problem systematically.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. *Proceedings of the 17th international conference on Computational linguistics*, 1:86–90.

Brian F Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological review*, 112(1):193.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and

C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: A multilingual metaphor repository. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 407–412. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Abhimanyu Dubey, Punit Jauhri, Abhinav Pandey, Abhishek Raghavan, Nikos Papaevagelou, Q. Gonzalez, and 1 others. 2024. The llama 3 herd of models. *arXiv*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Daniel Han and Unsloth Contributors. 2024. Unsloth: 2x faster llm fine-tuning. `https://github.com/unslothai/unsloth`. Open-source framework for efficient LLM fine-tuning.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. Lingualens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28220–28239.

Tina Krennmayr and Gerard Steen. 2017. Vu amsterdam metaphor corpus. In *Handbook of Linguistic Annotation*, pages 1053–1071. Springer Netherlands.

George Lakoff. 1993. The contemporary theory of metaphor.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. Framebert: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv*.

Michael Mohler, David Bracewell, David Hinote, and Marc Witbrock. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4219–4224. European Language Resources Association (ELRA).

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3170–3180.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pages 51575–51598.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. 2025. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*.

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. *arXiv*.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Lecture Notes in Computer Science*, pages 100–111.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *AAAI 31 (2017) 4444-4451*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and Amélie Héliou. 2024. Gemma: Open models based on gemini research and technology. *arXiv*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *Proceedings of the 8th Text REtrieval Conference*, volume 500, pages 77–82.

Hadi Wazni, Kin Ian Lo, and Mehrnoosh Sadrzadeh. 2024. Verbclip: Improving verb understanding in vision-language models with compositional structures. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 195–201.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and 1 others. 2024. Qwen2 technical report. *arXiv*.

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

A Appendix I: Ranks of select 10 metaphor-literal pairs

This section lists ranks of the literal words when retrieval attempt is made on a query metaphor. "H" stands for "hop".

A.1 Numberbatch

Table 13: Ranking Comparison across Direct and Hop Relations

Metaphor	Literal	Rank Direct	Rank H1	Rank H2	Rank H3
thunderbolt	motorcycle	251	234	555	854
cradle	sink	287	130	490	905
data	computer	4	2	29	26
tower	refrigerator	718	491	135	251
world	window	260	81	210	68
metal	car	494	89	83	101
broadcasting	tv	3	11	77	118
throne	sink	237	1283	1392	1437
alter	chair	1610	286	572	806
seat	chair	1	1	3	2

A.2 Gloves

Table 14: Comparative Ranking Analysis for Metaphorical Pairs

Metaphor	Literal	Rank Direct	Rank H1	Rank H2	Rank H3
thunderbolt	motorcycle	48	49	992	955
cradle	sink	700	52	648	1067
data	computer	19	1	27	33
tower	refrigerator	987	335	127	265
world	window	406	153	427	182
metal	car	136	182	126	181
broadcasting	tv	6	41	15	226
throne	sink	173	455	806	931
alter	chair	625	234	487	786
seat	chair	1	1	2	20
shell	bathtub	1451	69	254	232

A.3 Word2Vec

Table 15: Metaphorical Mapping and Hop-based Ranking Results

Metaphor	Literal	Rank Direct	Rank H1	Rank H2	Rank H3
thunderbolt	motorcycle	458	1138	694	913
cradle	sink	385	69	386	851
data	computer	5	1	33	50
tower	refrigerator	647	439	79	188
world	window	895	120	331	184
metal	car	442	138	126	153
broadcasting	tv	4	16	34	200
throne	sink	1670	1197	1571	1626
alter	chair	1441	158	362	667
seat	chair	1	1	1	2

B Appendix II: Selected Few Figurative Expressions

Table 16: Linguistic Expressions for Images in MS-COCO Dataset.

Original Caption:	The people are posing for a group photo.
Image:	
Synecdoche:	A sea of faces smiles for the camera.
Metaphor:	The people become a single frozen smile captured by the camera.
Euphemism:	The people are gathering together for a keepsake picture.
Personification:	The people gather together as if the camera has called them into a shared smile.
Politeness:	The individuals are politely posing together for a group photograph.
Epistemic:	It looks like the people might be gathering for what seems to be a group photo.
Original Caption:	A family standing next to each other while holding frisbees.
Image:	
Synecdoche:	A circle of hands grips frisbees as the family stands close together.
Metaphor:	The family forms a tight constellation, frisbees glowing in their hands.
Euphemism:	A family is spending a little quality time together with frisbees in hand.
Personification:	The family huddles close, the frisbees resting in their hands like eager companions.
Politeness:	A family is courteously standing together, each holding a frisbee.
Epistemic:	This appears to be a family who may be standing together, possibly holding frisbees.
Original Caption:	A long train traveling through a train yard.
Image:	
Synecdoche:	Iron wheels roll steadily through the train yard.
Metaphor:	The long train is a steel serpent winding through its yard.
Euphemism:	A lengthy train is making its way calmly through the rail area.
Personification:	The long train makes its way through the yard, stretching and sighing as it passes.
Politeness:	A long train is smoothly making its way through the train yard.
Epistemic:	It seems that a long train could be moving through what appears to be a train yard.
Original Caption:	A stop sign on the corner in front of a row of stores.
Image:	
Synecdoche:	A single red face watches the corner in front of the shops.
Metaphor:	The stop sign serves as a silent referee at the edge of the shops.
Euphemism:	A gentle reminder to pause sits at the corner before the line of shops.
Personification:	The stop sign stands at the corner, keeping watch over the line of waiting stores.
Politeness:	A stop sign is respectfully positioned at the corner in front of a row of stores.
Epistemic:	There seems to be a stop sign situated on a corner, perhaps in front of what looks like a row of stores.
Original Caption:	A man is grabbing the tie of another person that appears in the mirror.
Image:	
Synecdoche:	A grasp on the tie reflects a man confronting himself in the mirror.
Metaphor:	The man wrestles with his own reflection, tugging at a borrowed identity.
Euphemism:	A man is lightly adjusting another person's tie as seen in the mirror.
Personification:	The mirror seems to reveal a man reaching out, tugging at a tie as if it were alive.
Politeness:	A man is gently adjusting the tie of another individual as seen in the mirror.
Epistemic:	It appears that one man may be grabbing another's tie, as seen through a mirror reflection.