IRFL: Image Recognition of Figurative Language

Ron Yosef, Yonatan Bitton, Dafna Shahaf
The Hebrew University of Jerusalem
{ron.yosef, yonatan.bitton,dafna.shahaf}@mail.huji.ac.il

Abstract

Figures of speech such as metaphors, similes, and idioms allow language to be expressive, invoke emotion, and communicate abstract ideas that might otherwise be difficult to visualize. These figurative forms are often conveyed through multiple modes, such as text and images, and frequently appear in advertising, news, social media, etc. Understanding multimodal figurative language is an essential component of human communication, and it plays a significant role in our daily interactions. While humans can intuitively understand multimodal figurative language, this poses a challenging task for machines that requires the cognitive ability to map between domains, abstraction, commonsense, and profound language and cultural knowledge. In this work, we propose the Image Recognition of Figurative Language dataset to examine vision and language models' understanding of figurative language. We leverage human annotation and an automatic pipeline we created to generate a multimodal dataset and introduce two novel tasks as a benchmark for multimodal figurative understanding. We experiment with several baseline models and find that all perform substantially worse than humans. We hope our dataset and benchmark will drive the development of models that will better understand figurative language.

1 Introduction

Figures of speech include metaphors, similes, and idioms that allow language to be expressive, to convey abstract ideas that might otherwise be difficult to visualize, and to evoke emotion (Roberts and Kreuz, 1994; Fussell and Moss, 1998). A metaphor is a comparison between two unrelated concepts that enable us to think of the target concept in terms of the source concept. For example, in the sentence "You're a peach!", the person being addressed is equated with a peach, with the suggestion that the person is pleasing or delightful. A simile is a figure of speech that compares two things and is often introduced by "like" or "as" (Paul, 1970). A simile is called "open" when the shared properties are not explicitly revealed, like "Her heart is like a stone", and "closed" when they are explicitly revealed, like "Her heart is hard as stone". An idiom is a group of words with a figurative, non-literal meaning that can not be interpreted by looking at its individual words. For example, the idiom "We're on the same page" means "Agreeing about something (such as how things should be done)". Understanding metaphors and similes require the cognitive ability to map between domains, and depending on the source and target concept, it can require commonsense, association abilities, and general knowledge. Understanding

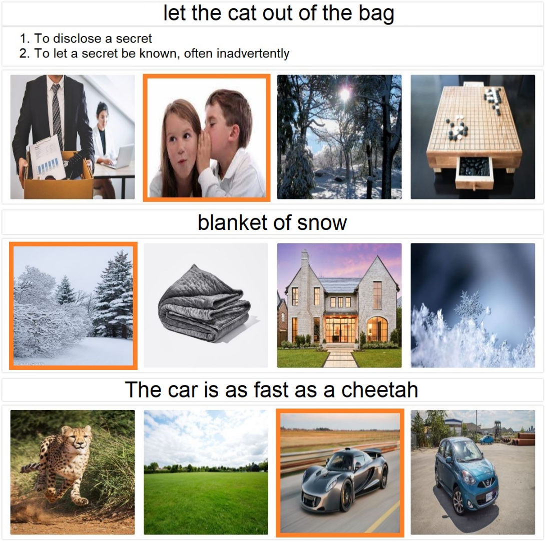

let the cat out of the bag
1. To disclose a secret
2. To let a secret be known, often inadvertently
blanket of snow
The car is as fast as a cheetah

Figure 1: Examples of the figurative understanding task for idiom, metaphor, and simile in corresponding order. The figurative phrase is displayed in the top section, and the bottom section displays four candidates from which the correct answer (orange) has been selected. Idiom tasks also display the idiom definitions below the idiom.

idioms requires profound language, and cultural knowledge (Paul, 1970; Philip, 2011). Humans intuitively understand these figures and employ them in everyday communication (Lakoff and Johnson, 1980; Hoffman and Kemper, 1987). These figurative forms are often conveyed through multiple modes, such as text and images, and frequently appear in advertising, news, social media, etc.

Due to its integral part in human communication, the detection and comprehension of multimodal figurative language is an important aspect of various multimodal challenges. Among these challenges are hate speech detection in memes (Das et al., 2020), fact-checking (Yao et al., 2022), sentiment analysis (Soleymani et al., 2017), humor recognition (Reyes et al., 2012; Schifanella et al., 2016), and identifying depression in social media posts (Yadav et al., 2020; Cheng and Chen, 2022). Figure 2 shows two photos posted on social media with metaphoric captions. In the left image, the caption reads, "Jumped off the sinking ship just in time", as this player left Chelsea - "the sinking ship", which is having a bad year, to join the leading team of the premier league, Arsenal. The right image was posted with the caption "A performing clown", as the person who is getting hit is a famous YouTuber who lost in a boxing match against a professional boxer. Multimodal figurative understanding is required to comprehend the metaphorical message being conveyed in these two posts. Vision and Language Pre-Trained Models' (VL-PTMs) understanding of figurative language combined with vision has not been thoroughly examined, if at all, partly due to the absence of large-scale datasets with ground truth labels of multimodal smilies, idioms, metaphors, etc.

In this work, we introduce the IRFL dataset of idioms, metaphors, and similes with matching figurative and literal images. We leveraged a textual dataset of idioms and an extensive pipeline we developed to find possible figurative and literal idiom images. We annotated these images via Amazon Mechanical Turk using the UI seen in (Appendix A.2) to create a large-scale dataset of idioms' figurative and literal images. In addition, we collected metaphors and similes' figurative and literal images. We then used the IRFL dataset to create two novel tasks of figurative understanding and figurative preference to examine the figurative understanding of Vision and Language models. The figurative understanding task evaluates VL-PTMs' ability to understand the relation between an image and a figurative phrase. The task is to choose the image that best visualizes the figurative phrase out of X candidates. Figure 1 shows an example of the task for idiom, metaphor, and simile. The preference task examines VL-PTMs' preference for figurative images. In this task, the model needs to rank figurative images of different categories correctly. Figure 3 shows the expected order versus the actual order of the idiom "ruffle someone's feathers" images based on the model scores. Finally, we experiment with generative models such as Dall-E and Stable Diffusion to examine their ability to generate figurative images for idioms.

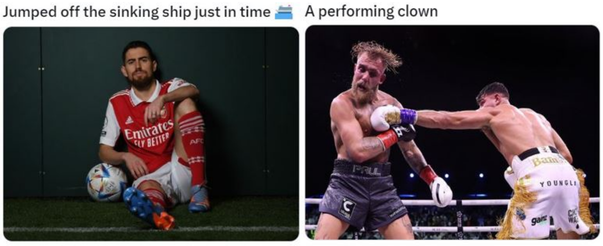


Figure 2: Two photos posted on social media. The left photo depicts football player Jorge Luiz Frello Filho Cavaliere wearing an Arsenal football club uniform. The right photo shows famous YouTuber Jake Paul taking a hit from professional boxer Tommy Fury during their boxing fight.

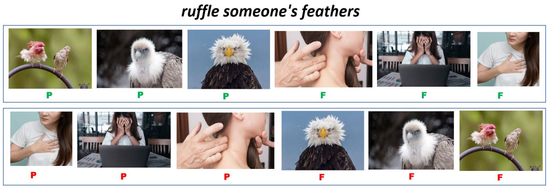


Figure 3: The Figurative and Partial Objects images of the idiom "ruffle someone's feathers" - "To unease, cause discomfort to someone" sorted from right to left by the CLIP-VIT-L/14 score. The images with the letter F are figurative, and the images with the letter P are partially literal. Green indicates correct rank, red indicates incorrect rank. The first row shows the ideal ranking order, while the second row shows the actual one. Figurative images with a green letter will appear in the first #F image from the right, and Partial Objects images with a green letter will appear in the last #P image. In this example, all of the models except LiT received 0 F_1 score.

2 The IRFL Dataset

Our goal is to introduce the IRFL dataset of idioms, metaphors, and similes with matching figurative and literal images and evaluate the figurative understanding and preference of Vision and Language models. To collect figurative and literal images for idioms, we developed an automatic pipeline that takes a list of idioms as input and outputs figurative and literal candidate images. We collected idioms from the MAGPIE corpus (Haagsma et al., 2020) of idiomatic expressions collected from Wiktionary, Oxford Dictionary of English Idioms, and UsingEnglish.com. The MAGPIE corpus contains 56,622 crowdsourced potentially idiomatic expressions, covering 1,756 unique idioms that appear in at least two of the dictionaries mentioned above. After collecting the idioms, we then feed them into the pipeline as input. First, we collect the definitions of the idioms from Wiktionary and Oxford dictionaries and construct search queries to find possible literal and figurative images (2.1). The process of choosing figurative and literal candidates involves several heuristics and implementation decisions elaborated at (2.2). We annotate the different relations between each idiom and its candidate images, thus creating the IRFL dataset (2.3). We evaluate the end2end dataset generation, and the fact that humans achieve high agreement helps to verify the correctness of the end2end process. The relation categories can be seen with a corresponding explanation in Table 1.

To collect metaphors and similes' images, we collected 35 textual metaphors and 142 textual similes from the internet. First, we constructed manual search queries and adapted the method used to search images in (2.2). Next, we annotated these images into "Figurative" and "Literal" categories. In total, we obtained 1107 figurative images and 1816 literal images for similes, and 333 figurative images and 729 literal images for metaphors. We verify the correctness of our dataset on different tasks in human evaluation section 3.3.1.

2.1 Search Queries

We want to find literal and figurative images for each idiom we collected from the MAGPIE dataset. For that, we collect the idioms' definitions from online dictionaries and parse them into "search queries". We first search Wiktionary for each idiom's definition and scrape the data using a web crawler. In case no definitions are found, we search the Oxford Dictionary and collect definitions using a similar method. The definitions in Wiktionary are usually tagged with the context in which they appear. For example, the idiom "white hat" has the "figurative" definition of "A good person; a hero", and the "slang" definitions of "a sailor" and "A well-meaning hacker". We collect this data and filter idioms with no "figurative" or "idiomatic" definitions. We then construct search queries by parsing the "figurative" and "idiomatic" definitions of idioms [2]. The parsing process separates definitions that are in fact several definitions concatenated into one. For example, we split the definition "A good person; a hero" into two search queries "A good person" and "A hero". In some rare cases, a definition may be an idiom, and to tackle such cases, we replace the idiom with its definitions.

Figurative Literal	The image conveys one or more definitions of the idiom to some extent, and it literally illustrates the phrase or visualizes the phrase objects/entities
Figurative	The image conveys one or more definitions of the idiom to some extent
Caption	The image illustrates the phrase literally
Partial Objects	The objects/entities of the phrase are visualized in the image
None	The image does not fit any of the categories

Table 1: Workers were guided to choose relation categories prioritized by the table's order. A scheme tree [1] was provided to illustrate how the correct thinking process should look like.

2.2 Choosing Images

To find figurative images for our search queries, we searched Google images [3], taking up to 20 images per search query. The resulting images included a lot of "garbage" and problematic images with specific characteristics, such as images in which the idiom they were derived from and its definitions are written. These images were problematic because a model may see a connection between an idiom and a figurative image solely based on the textual signal that appears in it. Such images were filtered

[1] https://irfl-dataset.github.io/assets/img/steps tree.PNG

[2] We also construct search queries from untagged definitions. Even though untagged definitions are rare (1-2% of all definitions), they are typically idiomatic.

[3] Images were searched with "SafeSearch" flag "on", and in "United States" region.

out by using OCR and a spelling tool to correct any spelling errors the OCR had. A large number of “garbage” images were found in the search results, including letters, postcards, newspapers, and images with mostly text in them. To tackle this problem, we used OCR to remove images with more than a couple of words and images with a text size bigger than 30%. In addition, we removed images that looked like documents that the OCR failed to detect. Images that passed these filters were literal, figurative, or had no connection to the phrase they originated from. Next, we calculated the matching score of each image with its phrase and search query. Images with a “phrase-image” score that passed a certain literal threshold (Appendix A.3) were tagged as “literal”, and from these images, we chose the top K images as literal candidates. From the non “literal” images, we chose the top K images with the highest “search query-image” score as Figurative candidates. We then annotated the relation between the figurative phrase and its Figurative and Literal candidates using the UI seen in Figure 4.

2.3 Human Annotation

We hired Amazon Mechanical Turk workers to annotate the relation between each idiom and its candidate images. Five workers annotated each image, the images were annotated in batches of five for the reward of $0.15 for batch. We created a difficult qualification test [4] to select quality annotators and provided them with an interactive training platform [5] to understand the task and the different categories better. We split the annotation process into batches with an average size of 60 idioms per batch. After each batch, We provided each worker with a personal profile page [6] to view its statistics and some handily picked examples where his choice was distant from a majority of four workers. We also provided workers with a leaderboard [7] that was updated after each batch to improve their competitiveness. Full annotation results and statistics are presented in Table 2.

The nature of this task is very subjective, and often the relation worker A sees between an idiom, and an image differs from the relation worker B see. We provide further discussion about this aspect of the task in (Appendix A.4). Despite the subjective aspect of the task and its complexity in distinguishing between the various categories, in 94% of the instances, there was a majority of 3 workers or more compared to a random chance of 29%. This shows that different people can see the same connection most of the time.

Categories	Figurative	Figurative Literal	Caption	Partial Objects	None	Total
Number	1970	751	434	487	2638	6697
3 workers majority	100%	100%	100%	100%	100%	94%
4 workers majority	75.5%	63%	68%	63%	80%	70%
5 workers majority	45%	33%	35%	38%	53%	43%
Mean per phrase	3.1	1.2	0.7	0.8	4	-
Median per phrase	2	0	0	0	4	-

Table 2: IRFL statistics on 628 idioms. The majority of the images have some relation to the figurative phrase. Most of the relations are Figurative.

3 Experiments

We evaluate humans and state-of-the-art image recognition models ability to understand figurative language (Section 3.3). We show that IRFL tasks are easy for humans (97% accuracy) and challenging for models (<27%). Additionally, we provide a detailed analysis per figure of speech, experiments with idioms and their definitions as input, and with different candidate types. We find that models fail the IRFL task due to their preference for partially literal images over figurative images and introduce a preference task to tackle this problem (Section 3.4). In addition, we examine the ability of generative models such as Dall-E and Stable Diffusion to generate figurative images for idioms (Section 3.5). We find that they are unable to generate figurative images given idiomatic phrases. Given the definitions of an idiom, generative models can generate figurative images.

3.1 Zero-Shot Baselines

We evaluate several diverse state-of-the-art vision-and-language models. Due to ViLT’s maximum sequence length of 40, we do not evaluate it on idioms. In all cases described below (except CLIP-ViL), the model encodes the figurative phrase and the image and produces a matching score for each pair. We chose the image that results the highest matching score as the image that best matches the figurative expression.

1. CLIP (Radford et al., 2021) is pre-trained

[4] https://irfl-dataset.github.io/mturk/image/qualification
[5] https://irfl-dataset.github.io/mturk/image/train
[6] https://irfl-dataset.github.io/profile/example
[7] https://irfl-dataset.github.io/mturk/leaderboard

with a contrastive objective that can be used without directly optimizing for the task. We use four versions of models with different amounts of parameters: RN50, ViT-B/32, ViT-L/14 and RN50x64/14 with 100M, 150M, 430M and 620M parameters respectively (RN50 was used during data collection).

2. CLIP-ViL (Shen et al., 2021), with 290M parameters, is a pre-trained vision-and-language model that uses CLIP as a visual backbone, rather than CNN based visual encoders that are trained on a small set of manually annotated data.

3. ViLT (Kim et al., 2021), with 111M parameters, incorporates text embeddings into a Vision Transformer (ViT).

3.2 Supervised Models

We join a line of benchmarks that introduce a test set without predefined train splits (Thrush et al., 2022; Rudinger et al., 2018; Emelin and Sennrich, 2021), (Bitton et al., 2022a). We believe that in order to understand metaphors and similes, a machine must be able to abstract and map between domains. It should be able to solve unseen cases without extensive training (Mitchell, 2021). Contrary to metaphors and similes, understanding idioms requires language and cultural knowledge that can be learned through extensive training. We train a supervised model for figurative classification of idioms. We add a binary classifier on top of the pre-trained embeddings to classify whether a given image is figurative or not. We use CLIP (VIT-B/32) model, concatenate the textual idiom embedding to the visual image embedding, followed by a classifier that produces a matching score, where a matching score above 0.5 is labeled 'Figurative'. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001, batch size of 12, and train for 7 epochs. We run the fine-tuned model on the understanding and preference task using the model's matching score. We train the binary classifier on 4790 images for the understanding task and 3802 images for the preference task[8]. We repeat five experiments with different random seeds for each task and take the mean score along with the standard deviation.

3.3 Understanding Task

The figurative understanding task evaluates VL-PTMs' ability to understand the relation between an image and a figurative phrase. The task is to choose the image that best visualizes the figurative phrase out of X candidates. Our goal was to create an understanding task that consists of "mixed" candidate types and represents the richness of our dataset. The "mixed" tasks provide a holistic image of the figurative understanding of vision and language models. We constructed and crowdsourced 810 "mixed" figurative understanding task instances for idioms, metaphors, and similes.

The basic structure of all "mixed" instances is the same. Each instance contains four candidates, of which one is the correct answer, and $1-3$ candidates are partially literal distractors. The "mixed" idiom instances have one to two partially literal distractors and one to two random images. The simile "mixed" instances contain a distractor image of the target concept without the compared property or with its antonym visualized, a distractor image of the source concept, and one random image. The metaphor "mixed" instances consist of between one to three partially literal distractors, and the remaining candidates are random images. In 65% of the idioms understanding task instances, the correct answer is "Figurative", and in the other 35%, the correct answer is "Figurative Literal". Figure 1 shows two examples of the "mixed" figurative understanding task for metaphor and simile.

3.3.1 Human Evaluation

We asked annotators that did not work on previous IRFL tasks to solve the figurative understanding task. Each instance of the "mixed" understanding task was annotated by 5 annotators, and the correct answer is chosen by the majority. We find that human performance on the data sampled for all figures of speech ranges between $90\% - 100\%$. Additionally, in $83\% - 99\%$ of the instances, there was an agreement between at least four annotators compared to a random chance of 6%. Samples from the validation process are presented in Appendix A.5.

3.3.2 Results and Model Analysis

Zero-shot results on the "mixed" figurative understanding task are presented in Table 3. The

[8] Training data does not contain any of the images or idioms that appear in the task.

Categories	Idioms		Metaphors	Similes
	Figurative	Figurative Literal		
Humans	97%	90%	99.7%	100%
CLIP-VIT-L/14	17%	**56%**	25%	**52%**
CLIP-VIT-B/32	16%	44%	23%	45%
CLIP-RN50	14%	37%	27%	47%
CLIP-RN50x64/14	22%	**56%**	**30%**	52%
LiT	**27%**	31%	21%	19%
ViLT	-	-	23%	40%
# Unique Phrases	48	30	35	142
# Tasks	135	65	333	277

Table 3: Zero-shot models performance on the IRFL "mixed" understanding task by figurative type. There are two columns for idioms, the first column represents the score for the "Figurative" images, and the second for the "Figurative Literal" images. In the idioms tasks the model received both the Idioms and their definitions as input. Numbers are the percentage of instances annotated correctly. Bold numbers indicate the best model performances.

best model achieved 27%, 30%, and 52% accuracy on the idioms[9], metaphors, and similes tasks compared to a random chance of 25%. **These results suggest that models do not understand the connection between a figurative phrase and an image like humans do.** We conduct a fine-grained analysis to examine if models failed the "mixed" understanding task because they do not see any connection to the figurative images or rather because they prioritize "weak" literal connections over figurative ones.

Models prefer partially literal images over figurative ones. We analyzed the models' choices on the "mixed" figurative understanding task and found that in all models (excluding LiT on idioms and similes), a partially literal distractor was selected in 92% − 100% of the instances where the models failed across all figures of speech (idioms, metaphors, and similes) . This shows that models prefer partially literal images over figurative ones. **We find the case of idioms to be particularly interesting in this regard. Models receive a relatively long prompt containing both the idiom and its definitions as input. Instead of picking an image that fits the prompt semantically, they choose an image that is literal to one or two words.**

Models partially understand the figurative connection between idioms and images. To examine whether models can comprehend a figurative connection between an image and an idiom, we experiment with random candidates and several configurations of the understanding task (Table 4). The accuracy score on the Figurative category with 2 candidates is 61% − 87%, and 22% − 71% with 4 candidates. These results are marginally above random chance but still below human performance on the "mixed" task. When given the idiom alone as input, most models achieved 80% − 84% with 2 candidates and 30% − 46% with 4 candidates compared to random chance of 50% and 25%. These results suggest that models partially understand the figurative connection between idioms and images. Moreover, we see a significant performance drop with all models when increasing the number of candidates.

Categories	Figurative						Figurative Literal	
Candidates	2		4		6		4	
Random	50%		25%		16.6%		25%	
CLIP-VIT-L/14	64%	**87%**	**46%**	**71%**	**33%**	**53%**	76%	**86%**
CLIP-VIT-B/32	61%	84%	38%	67%	30%	**53%**	65%	82%
CLIP-RN50	56%	75%	30%	60%	24%	46%	**78%**	**86%**
CLIP-RN50x64/14	**67%**	79%	38%	67%	27%	51%	69%	85%
LiT	57%	61%	22%	22%	19%	18%	17%	24%

Table 4: Zero-shot models performance on the different configurations of the idiom understanding task with random candidates. Numbers are the percentage of instances annotated correctly. Bold numbers indicate the best model performance. There are three configurations for the Figurative Category with 2, 4, and 6 candidates. The table is double-columned. The left column shows the score for the phrase alone as input, and the right column shows the score for the phrase and its definitions as input.

In the Figurative Literal category, models achieve a 65% − 78% accuracy score with 4 candidates, significantly higher than the performance in the Figurative category with 2 and 4 candidates. These results can be explained by the fact that Figurative Literal images possess a literal connection to the phrase in addition to a figurative one.

Models understand metaphors, but fail to reach human performance. Table 5 shows the models' performance on the metaphors figurative understanding task with random candidates. The accuracy score of all models, excluding LiT, on the Figurative category with 2 candidates is 72% − 88%, and 53% − 76% with 4 candidates. We see a significant performance drop with all

[9] Idioms were passed along with their definitions as input.

Categories	Metaphors		Similes	
Candidates	2	4	2	4
CLIP-VIT-L/14	87%	72%	**99%**	**97%**
CLIP-VIT-B/32	86%	73%	**99%**	**97%**
CLIP-RN50	83%	66%	**99%**	**97%**
CLIP-RN50x64/14	**88%**	**76%**	98%	96
LiT	47%	27%	49%	24%
ViLT	72%	53%	96%	91%

Table 5: Zero-shot models performance on the metaphors and similes understanding task with random candidates. Numbers are the percentage of instances annotated correctly.

models when increasing the number of candidates. The results suggest that models understand metaphors but fail to reach human performance.

Models understand similes as well as humans. Table 5 shows the models' performance on the similes figurative understanding task with random candidates. The accuracy score of all models, excluding LiT, on the Figurative category with 2 candidates is $96\% - 99\%$, and $91\% - 97\%$ with 4 candidates. Models' performance is competitive with that of humans, and the models maintain their performance when increasing the number of candidates. We note that we experiment with open similes where the compared property is explicitly mentioned in the simile. Thus the Figurative images can be seen as Figurative Literal. As we analyzed the "mixed" understanding task results in more depth, we found that across all models excluding LiT, $55\% - 61\%$ of the figurative images received a higher matching score than the source concept images. In addition, $50\% - 66\%$ of the source concept images received a higher matching score than the target concept distractor image. These suggest that models prioritize simile images in the following order: 1) images of the target concept with the compared property, 2) images of the source concept, 3) images of the target concept without the compared property.

Fine-tuning improves figurative understanding and reduces partially literal preference. Fine-tuning results are presented in Table 6. The mean Figurative category accuracy is 58% compared to 13% in the Zero-shot configuration. We analyzed the fine-tuned model results and compared them to the zero-shot configuration and found that in $41\% \pm 4.3$ of the instances where the model failed, a partially literal distractor was selected compared to 96% in the zero-shot configuration. Along with this improvement in literal preference, Figurative Literal category accuracy raised from 41% in zero-shot to 49%. These results show that models can moderate their preference for partially literal images and recognize idiomatic figurative connections better, using extensive training. Moreover, the results suggest that the data is a valuable training signal for this task.

Categories	Figurative	Figurative Literal
Zero-Shot	16%	41%
Supervised	$58\% \pm 4.2$	$49\% \pm 2.6$

Table 6: Supervised models performance. Results are the mean and standard deviation of the accuracy of five experiments.

3.4 Ranking Task Analysis

To tackle vision and language models' strong preference toward partially literal images over figurative images, we introduce the preference task. The preference task is to rank the Figurative images higher than partially literal distractors based on the model matching score. First, we rank the figurative phrase images by their matching score from higher to lower, then we define two classes, $\#F$ which consists of the Figurative images, and $\#P$ which consists of the partially literal images. The model then predicts the first $\#F$ images as Figurative and the last $\#P$ images as partially literal images, the F_1 score of the model predictions is the preference task score. The results of the preference task are presented in Table 7. We evaluate all figurative phrases that have images from both of the categories. Models' scores on the

Ranking	Idioms		Metaphors	Similes
	Figurative Literal	Figurative		
CLIP-VIT-L/14	57	37	26	**44**
CLIP-VIT-B/32	54	36	22	38
CLIP-RN50	54	37	25	38
CLIP-RN50x64/14	**61**	39	**29**	43
LiT	54	**56**	25	25
ViLT	-	-	23	34
# of phrases	94	149	35	142

Table 7: The preference task performance, the scoring metric is F_1. The Idiom category is double-columned. The left column shows the score for Figurative Literal images, and the right column shows the score for Figurative images.

preference task are low (<61%). We expect models with proper figurative preference to achieve better results. Models' success in the Figurative Literal category can be attributed to the literal connections of the Figurative Literal images.

Categories	Figurative	Figurative Literal
Zero-Shot	36	54
Supervised	68 ± 3.8	64 ± 2.25

Table 8: Supervised models performance. Results are the mean and standard deviation of the F_1 score of five experiments.

The supervised model, after fine-tuning, achieved a 68 ± 3.8 F_1 score on the Figurative category, almost double the zero-shot score of CLIP-ViT-B/32 (36). Additionally, the score in the Figurative Literal category was improved by 10 ± 2.25 points. These results align well with the observation that the fined-tuned understanding task model showed substantially moderate literal preference. Table 8 shows the fine-tuned model results.

3.5 Generative Models Analysis

To examine whether generative models can generate figurative images, we sampled 15 idioms from the IRFL dataset and experimented with the idioms and their definitions as input to Dall E and Stable Diffusion. We annotated 345 generated images and found that generative models failed to generate figurative images for given idioms but instead generated literal images. When provided with the definitions as input, the models succeeded in creating figurative images to some extent. Statistics on the generated images and the matching IRFL images can be seen in Table 9.

Categories	Dall E		Stable Diffusion		IRFL	
Figurative	0%	42.5%	0%	11%	4%	46%
Figurative Literal	0%	10%	5%	1%	20%	6%
Caption	31%	0%	17%	0%	35%	0%
Partial Objects	48%	2%	42%	2.5%	23%	1.5%
None	19%	44%	27%	85%	4%	43%
Number	48	120	59	118	69	126

Table 9: The table is double-columned, the first column describes the percentage of images generated by idioms, and the second column describes the percentage of images generated by the idioms' definitions.

The experiment results show that our pipeline extracted more Figurative, Figurative literal, and Caption images and fewer None images than the generative models. Future work might focus on the quality of generative models' figurative images and the emotions they evoke.

4 Related Work

4.1 Commonsense

Common sense is a topic of increasing interest [36]. Many commonsense reasoning tasks have been proposed, both in NLP (Zellers et al., 2019b; Sap et al., 2019a; Forbes et al., 2019; Saha et al., 2021), and computer vision (Marino et al., 2019; Zellers et al., 2019a; Park et al., 2020; Bitton-Guetta et al., 2023) ranging from physical context (Bisk et al., 2020) to social interactions (Sap et al., 2019b). A particularly relevant line of work are abstractions (Ji et al., 2022), associations (Bitton et al., 2022a), and analogies (Bitton et al., 2022b): understanding metaphors and similes often require association, abstraction, and general knowledge depending on the target and source concepts and the metaphorical message. For example, understanding the simile "as stubborn as a mule" requires the common sense knowledge that mules are stubborn (where in fact, they are not). The metaphor "John is a fox" uses the association of foxes with slyness.

4.2 Idioms

Previous work on idioms focused on the detection, interpretation, and representation of textual idioms (Fazly et al., 2009; Verma and Vuppuluri, 2015; Peng and Feldman, 2016; Salton et al., 2016; Liu and Hwa, 2017; Li and Sporleder, 2009; Liu et al., 2017; Liu and Hwa, 2016; Zhou et al., 2021). Recently, several papers have examined the ability of pre-trained LMs to represent idioms. Shwartz and Dagan (Shwartz and Dagan, 2019) found that LMs' representation of idiomatic expressions was of lower quality than that of literal ones. Chakrabarty at el. (Chakrabarty et al., 2022) introduced a narrative understanding benchmark focused on interpreting figurative language and found that pre-trained LMs irrespective of their size, struggle to perform well in zero-shot and few-shot settings. However, to the best of our knoweldge, Vision and Language Pre-trained models (VL-PTMs) understanding of idioms has not been investigated until this work.

4.3 Metaphors and Similes

Metaphors and similes have been studied previously primarily in terms of interpretation, generation, and detection of textual metaphors and similes (Aghazadeh et al., 2022; Stowe et al., 2021; He et al., 2022; Zeng et al., 2019; Chakrabarty et al., 2020, 2022). Recently there have been several works focusing on the ability of VL-PTMs to understand similes and metaphors. Zhang et al. (Zhang et al., 2021) introduced the first large-scale multimodal dataset of metaphors. Liu and Giegle (hen Liu et al., 2022) presented FigMemes, a dataset for figurative language classification in politically-opinionated memes. Akula et al. (Akula et al., 2022) annotated a visual advertisement dataset with similes as captions to introduce MetaCLUE, a set of vision tasks on visual metaphor. We find MetaCLUE the closest to ours concerning similes. The key difference between IRFL and MetaCLUE is the tasks and images. MetaCLUE's images are synthetic, while ours are more natural. Additionally, our tasks introduce a new aspect of preference (literal vs. figurative) into multimodal metaphorical understanding.

5 Limitations and Conclusions

We introduced IRFL, a dataset of Figurative and Literal images for idioms, metaphors, and similes. We introduced two novel tasks as a benchmark for figurative understanding. Our tasks are easy for humans and challenging for state-of-the-art models. We provided an extensive evaluation of the dataset.

Our pipeline has several limitations. Future work can focus on improving the pipeline, in particular, improving the quality of figurative candidates for idioms, and increasing the automation for metaphors and similes. We hope that the IRFL dataset and benchmark will drive the development of models that will better understand figurative language.

6 Acknowledgements

We would like to thank Nir Sweed for his valuable feedback. We would also like to thank Nissim Barzilay for his contribution to the collection of figurative and literal images for metaphors and similes. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant no. 852686, SIAM, Shahaf).

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2022. Metaclue: Towards comprehensive visual metaphors research.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. 2022a. Winogavil: Gamified association benchmark to challenge vision-and-language models. *ArXiv*, abs/2207.12576.

Yonatan Bitton, Ron Yosef, Eli Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. 2022b. Vasr: Visual analogies of situation recognition. *ArXiv*, abs/2212.04542.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *arXiv preprint arXiv:2303.07274*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.

Ju Chun Cheng and Arbee L. P. Chen. 2022. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *CoRR*, abs/2012.14891.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *ArXiv preprint*, abs/1908.02899.

Susan R Fussell and Mallie M Moss. 1998. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Qi He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *ACL*.

hen 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*, page 7069–7086. Association for Computational Linguistics.

Robert R. Hoffman and Susan Kemper. 1987. What could reaction-time studies be telling us about metaphor comprehension? *Metaphor and Symbol*, 2:149–186.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D. Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The Journal of Philosophy*, 77(8):453–486.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323, Singapore. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3230–3236. AAAI Press.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. *CoRR*, abs/1906.00067.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*.

Anthony M. Paul. 1970. Figurative language. *Philosophy and Rhetoric*, 3(4):225–248.

Jing Peng and Anna Feldman. 2016. Experiments in idiom recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2752–2761, Osaka, Japan. The COLING 2016 Organizing Committee.

G. Philip. 2011. *Colouring Meaning: Collocation and connotation in figurative language*. Studies in Corpus Linguistics. John Benjamins Publishing Company.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*, 74:1–12. Applications of Natural Language to Information Systems.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. *CoRR*, abs/1608.02289.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *ArXiv preprint*, abs/2107.06383.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. 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 6724–6736, Online. Association for Computational Linguistics.

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. *ArXiv preprint*, abs/2204.03162.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 681–687, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2019. Neural simile recognition with cyclic multitask learning and local attention. In *AAAI Conference on Artificial Intelligence*.

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, Online. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Appendix

A.1 Dataset Supplementary Materials

1. Dataset documentation, metadata, and download instructions are available at `https://irfl-dataset.github.io/download`.

2. Intended uses: We hope researchers will use our benchmarks to evaluate Vision and Language models. We also hope that our pipeline and dataset will inspire future work on creating extensive multimodal datasets of other figures of speech.

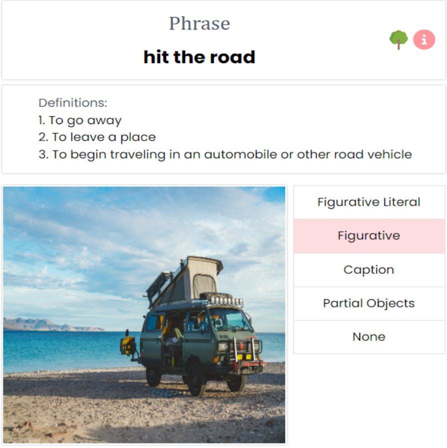
3. Author statement: We bear all responsibility in case of violation of rights in using our benchmark.

4. Licenses: Code is licensed under the MIT license `https://opensource.org/licenses/MIT`. Dataset is licensed under CC-BY 4.0 license `https://creativecommons.org/licenses/by/4.0/legalcode`.

5. Hosting & preservation: our website is deployed, and all data is accessible and available. We encourage researchers to send us model predictions on the created test sets. We will update a model and players leaderboard with these results periodically.

6. Code repository: `https://github.com/irfl-dataset/irfl`

A.2 Annotation UI



Phrase

hit the road

Definitions:
1. To go away
2. To leave a place
3. To begin traveling in an automobile or other road vehicle

Figure 4: The UI used to annotate the automatic pipeline candidate images. Annotators need to choose the category that best describes the relationship between the idiom and the image.

A.3 Literal Threshold

To find a literal threshold, we conducted two grid searches on images that passed the OCR filters and had a "phrase-image" score higher than the "search-query" score. We sampled 20 images from each point in the distribution of $-10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10$, and annotated them as "literal" or "non-literal". This distribution aligns with the normal distribution of the images that stand the two criteria mentioned above (Figure 5). We found the range of $(-2, 2)$ to result in the best thresholds, and so we conducted a more dense grid search in this range. We sampled 30 images from each point in the distribution of $-5, -4, -2, -1, 0, 1, 2, 4, 5$, and annotated them as "literal" or "non-literal". We chose the threshold of 1.150353 with a TPR of 86% and FPR of 18%.

We observed that when the "phrase-image" score is high, we can say that the image is literal with a high probability. However, the reverse is not

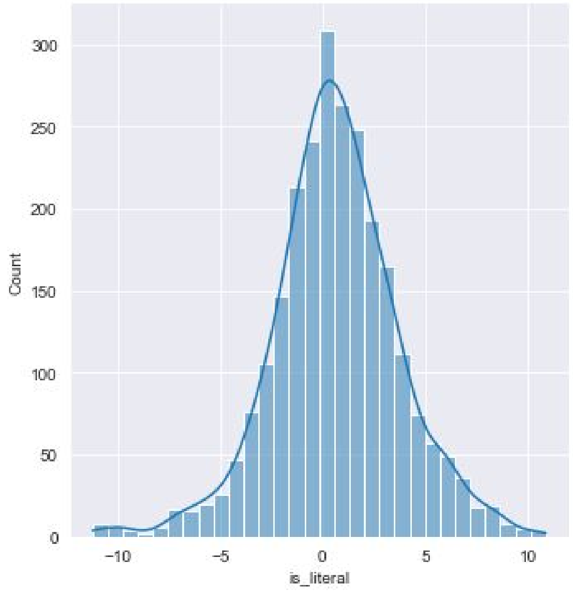


Figure 5: The distribution of the images that passed the OCR filters and had a "phrase-image" score higher than the "search-query" score.

true, there can be multiple “literal” images with a very low literal score (Figure 6).

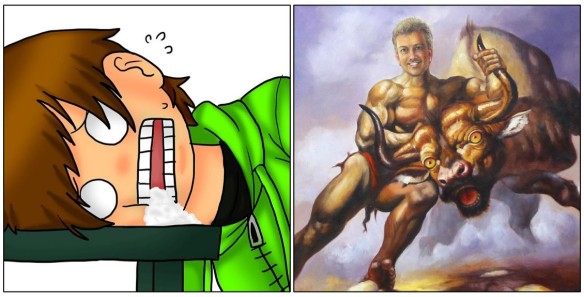
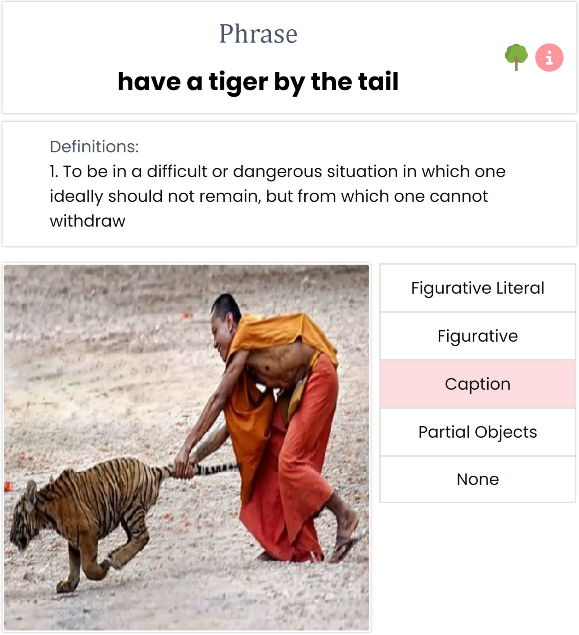


Figure 6: Literal images of the idiom "Foam at the mouth" and the idiom "Take the bull by the horns". Both images have a "phrase-image" score of -9.

A.4 Annotation Task Discussion

The nature of the image annotation task is very subjective, and often the relation worker A sees between an idiom and an image differs from the relation worker B sees. The connection a worker sees between an image and an idiom can vary based on his understanding of the image scene and his interpretation of that scene. For example, Figure 7 shows a "Caption" image of a person holding a tiger by the tail in a non-dangerous or difficult situation in which one should not remain. The person is smiling as it seems like he is playing with a very young tiger (small in size). The majority of workers agreed with this explanation, except one who disagreed and chose the "Figurative Literal" category. This worker’s interpretation arose from her viewpoint as a mother, and she said, "as a mother, I would still say that it’s dangerous and the person is being foolish".



Phrase

have a tiger by the tail

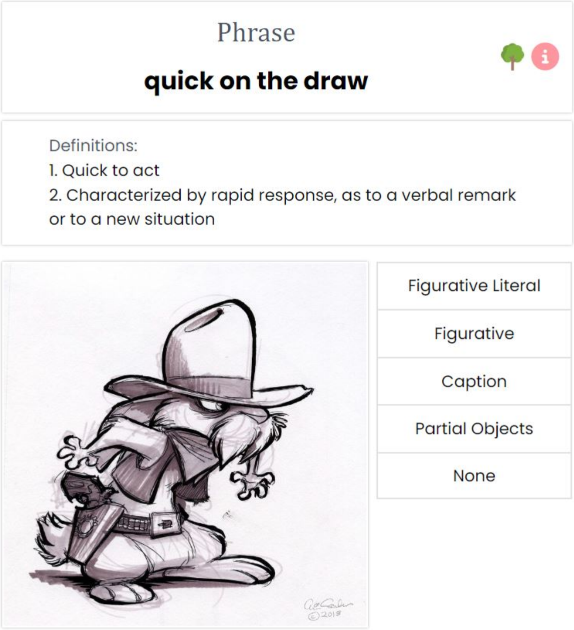
Definitions:

1. To be in a difficult or dangerous situation in which one ideally should not remain, but from which one cannot withdraw

Figure 7: A candidate image from the training platform.

Another example of different interpretations is the image seen in Figure 8, which shows an image of a cowboy bunny drawing with the idiom - "quick on the draw". The annotations of this image were very diverse as 5 workers chose 4 different categories. Two workers chose "Figurative" as they saw a connection to the idiom definitions. One worker chose "Figurative Literal" as he saw a connection to the idiom definitions and a literal connection to the idiom. Another worker chose "Caption" because he did not find the image to be "Figurative" and saw the idiom literally as illustrating the image. The last worker selected "None" as he did not find a clear literal connection and didn’t see the image as "Figurative" as it lacked an indication that the bunny was "quick to act" or "characterized by rapid response".

A.5 Understanding Task Samples

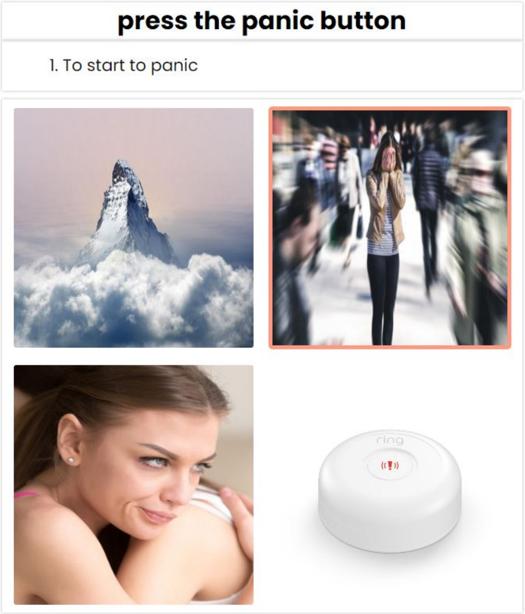


Phrase

quick on the draw

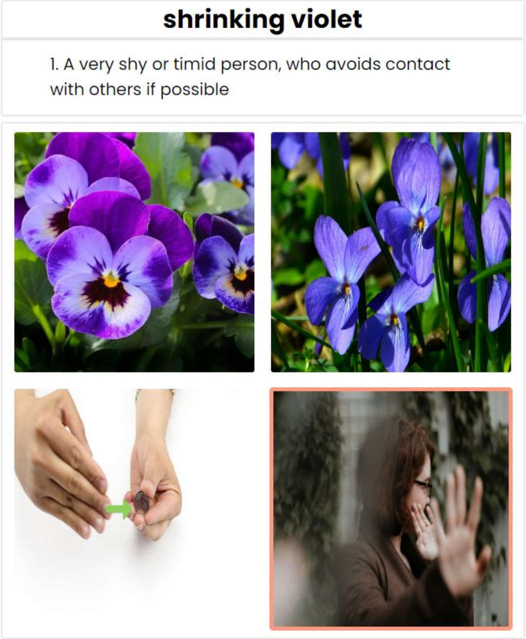
Definitions:
1. Quick to act
2. Characterized by rapid response, as to a verbal remark or to a new situation

Figure 8: A figurative candidate of the idiom "quick on the draw".



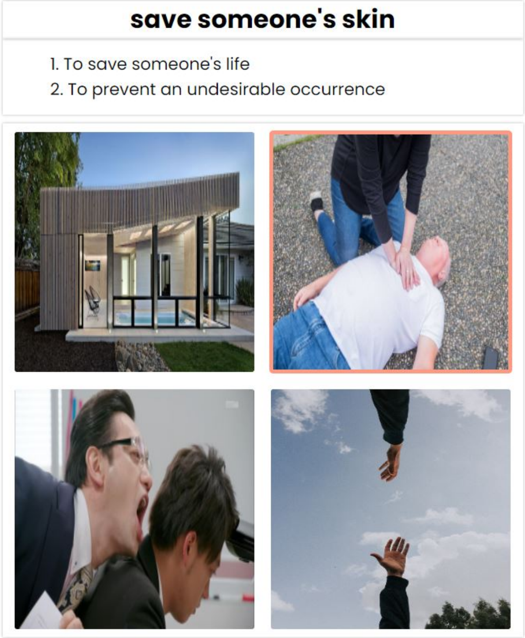
press the panic button

1. To start to panic



shrinking violet

1. A very shy or timid person, who avoids contact with others if possible

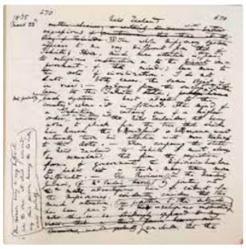


save someone's skin

1. To save someone's life
2. To prevent an undesirable occurrence

get the boot

1. To be dismissed from employment
2. To be voted out, evicted, or otherwise made to leave

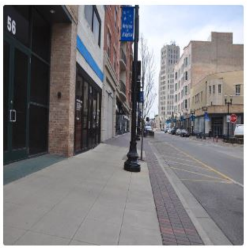


heart of gold

a night owl

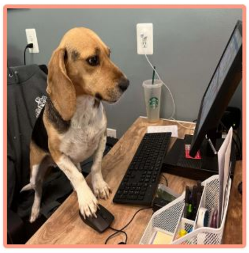
the car is a rocket

jungle city

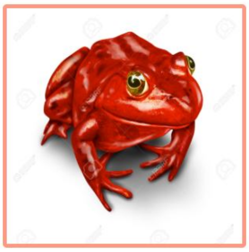


The juice is as sweet as sugar

The dog is as busy as a bee



The frog is as red as a tomato



The milk is as fresh as a daisy