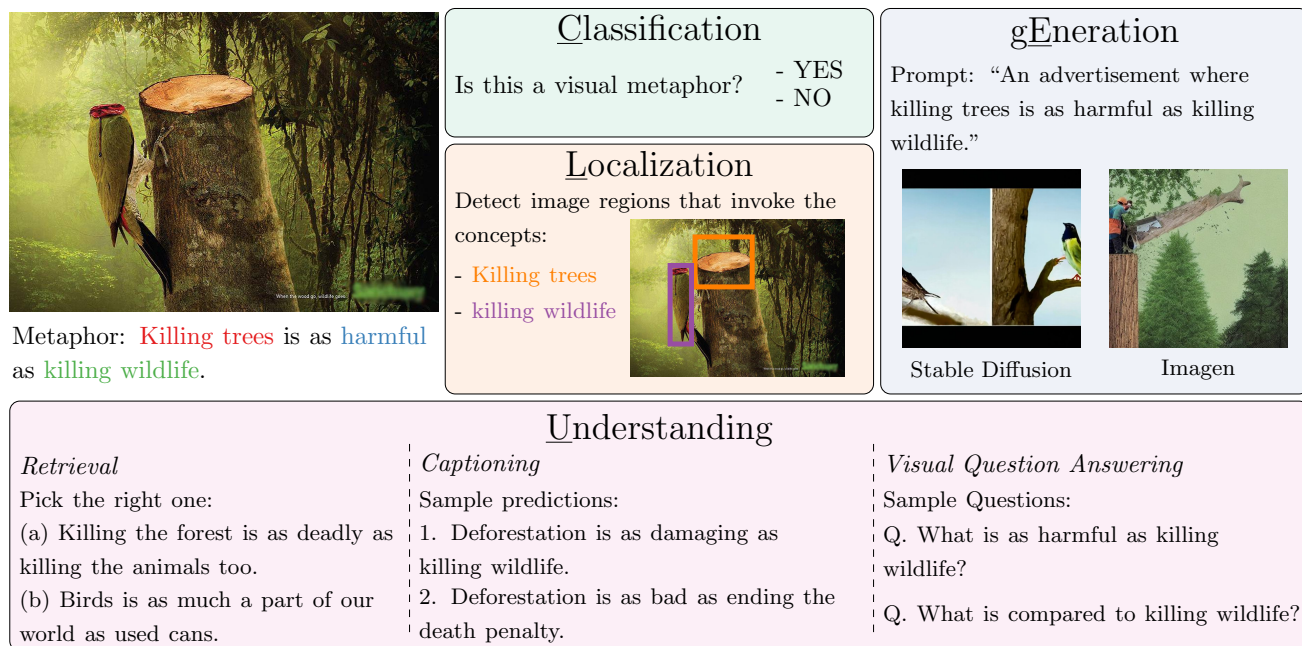# MetaCLUE: Towards Comprehensive Visual Metaphors Research

Arjun R. Akula*, Brendan Driscoll*, Pradyumna Narayana, Soravit Changpinyo,
Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu,
Leonidas Guibas, William T. Freeman, Yuanzhen Li, Varun Jampani*

Google

Figure 1. With **MetaCLUE**, we introduce several interesting tasks related to visual metaphors. We collect metaphor annotations (objects, abstract concepts, relationships and object boxes) for evaluating existing models on these tasks. Specifically we perform a comprehensive evaluation of vision and language models on four different tasks (**C**lassification, **L**ocalization, **U**nderstanding, and g**E**neration). Comprehensive experiments in this work show that state-of-the-art techniques mostly focus on literal interpretation and perform poorly in understanding and generation of metaphor images.

Abstract

Creativity is an indispensable part of human cognition and also an inherent part of how we make sense of the world. Metaphorical abstraction is fundamental in communicating creative ideas through nuanced relationships between abstract concepts such as feelings. While computer vision benchmarks and approaches predominantly focus on understanding and generating literal interpretations of images, metaphorical comprehension of images remains relatively unexplored. Towards this goal, we introduce MetaCLUE, a set of vision tasks on visual metaphor. We also collect high-quality and rich metaphor annotations (abstract objects, concepts, relationships along with their corresponding object boxes) as there do not exist any datasets that facilitate the evaluation of these tasks. We perform a comprehensive analysis of state-of-the-art models in vision and language based on our annotations, highlighting strengths and weaknesses of current approaches in visual metaphor classification, localization, understanding (retrieval, question answering, captioning) and generation (text-to-image synthesis) tasks. We hope this work provides a concrete step towards developing AI systems with human-like creative capabilities. Project page: https://metaclue.github.io

---

*Equal Contribution

1. Introduction

"Metaphor is pervasive in everyday life ... Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature." — Lakoff & Johnson [25]

Creativity is a process of generating a new perspective on a problem or a situation. Metaphorical thinking has been recognized as a key and powerful mechanism of creativity [24, 29, 50]. Humans engage metaphors in their creative thinking process as strategies to link or blend concepts, or to view a concept from a target domain in terms of another, apparently dissimilar concept from a source domain [25]. Metaphors also provide a sophisticated tool for nuanced human communication. Let us take a closer look at the structure of metaphors – and especially visual metaphors.

Metaphors[1] are a cognitive construct in which a concept is compared to a seemingly unrelated concept via some shared attribute. Take as an example 'This car is a cheetah', where 'This car' is compared to 'a cheetah' in terms of speed. Metaphors have a simple syntactic structure of 'A is B' where A is referred to as the **primary concept** and B as the **secondary concept**. The implied analogy in a metaphor is of the form: '(primary concept) is as (relationship)[2] as (secondary concept)' and often involves an attribute transfer from the secondary to the primary concept. Some examples include 'This phone is as fast as a rocket', 'Cigarettes are as harmful as bullets' etc. The primary and secondary concepts are usually unrelated at a glance, resulting in an element of surprise and creativity in metaphorical expressions. Despite following such simple structure, metaphors are quite powerful in conveying creative ideas. Metaphors are pervasive in all forms of communication, such as speech, text, visual etc.

Visual Metaphors are images where the primary and secondary concepts are visually depicted in an image conveying the metaphorical message to the viewers. Visual metaphors are widely used in mass media communications like advertising and journalism [15, 43, 45]. In this work, we work with Ad images, as metaphors tend to be prevalent in ads. There are numerous ways a metaphor can be represented visually. Following the classification in [15], there are at least 4 different types of visual metaphors. Fig. 2 shows sample images that belong to these types along with our annotations of primary, secondary concepts and their relationship. In *contextual* metaphors, either the primary or secondary concept is not explicitly visible, but is inferred from the context (e.g., apple in the left-most image). In *hybrid* metaphors, the primary and secondary concepts are

visually conflated. *Juxtaposition* forms one of the simplest visual metaphor types, where the two concepts are just presented next to each other. *Multimodal* metaphors represent one of the concepts with another modality, such as text or logo. In practice, visual metaphors use multiple of these strategies to convey a metaphor in an effective manner. In many cases, the implied metaphorical meaning is somewhat open-ended. Interpretation of visual metaphors depends on several external factors, such as familiarity with the brands and cultural context.

These visual variations and nuances make automatic cognition or generation of visual metaphors highly challenging. While the last decade has seen rapid progress in many areas of understanding and generation tasks, prior works in computer vision focus heavily on literal interpretation of images and overlook the importance of metaphorical reasoning in understanding the image message [1, 48]. We believe that developing AI systems with metaphorical comprehension and generation capabilities can greatly assist humans in creative endeavors involving conveying concepts in new and exciting ways. Such systems provide an important step towards conferring human-like creativity to AI models.

To this end, we introduce multiple interesting tasks and construct metaphor annotations that enable comprehensive research on visual metaphors. As metaphors are more common in visual Ads, we start with the Pitt's Ads dataset images [20] and then perform a rigorous multi-stage annotation process with expert annotators to filter metaphorical images, add metaphor annotations, and perform additional validation steps to clean the annotations[3]. While there is recent work making advances in understanding non-literal interpretations in natural language research [7, 12], this work proposes the first step towards metaphor analysis in images.

As illustrated in Fig. 1, we perform comprehensive evaluations with state-of-the-art techniques on four sets of tasks, which we call **MetaCLUE**: 1. C̲lassification: This is binary classification task of estimating whether a given image contains a metaphor or not. In other words, *Are visual features indicative of whether there exists a metaphor in a given image or not?*. 2. L̲ocalization: Here, the task is to *localize the image regions that invoke the primary and secondary concepts in the viewer.* This is similar to a standard object detection task, but is more complicated in the case of visual metaphors as the primary/secondary concepts may not be explicitly present in an image. 3. U̲nderstanding: *Can our models understand the metaphorical message in a given image?* We pose this understanding problem as 3 tasks where we can quantitatively measure the performance: Retrieval, Captioning and Visual question answering. 4. g̲Eneration: *Can we generate an image that conveys the metaphor, given the metaphorical message as a text prompt?*

---

[1]Grammarians distinguish a metaphor "A is B" from a simile "A is like B". In our work we use "metaphor" to encompass both variants.

[2]We use the word 'relationship' to denote the shared property of primary and secondary concepts, usually adjectives or adjectival phrases.

[3]The Ads, while useful for the purposes of our paper, some images may perpetuate harmful stereotypes according to characteristics such as gender.

Contextual

This beer is as tasty as a real apple.

This car is as adventurous as a space ship.

Hybrid

Driving this SUV is as smooth as birds flying in the sky.

This pencil is as red as a fire truck.

Juxtaposition

This chocolate bar is as rich as gold.

This car is as made for the beach as a crab.

Multimodal

These donuts are as unique as as talking people.

The car is as rugged as this muddy trailer.

Figure 2. **Sample Visual Metaphors with their Annotations.** There are different types of visual metaphors. The type depends on how the primary and secondary concepts are visually depicted. Here are sample Ad images from [20] where we annotated the primary concept, secondary concept and their relationship.

We comprehensively evaluate existing state-of-the-art techniques for each of these tasks on our collected metaphor annotations. We evaluate the models both in a zero-shot manner as well as with finetuning on our annotations. Even though finetuning resulted in some improvements, most models struggle to produce satisfactory results in many cases, demonstrating the difficulty of these tasks. Our experiments highlight several strengths and weaknesses of the existing techniques on comprehending and generating visual metaphors, providing a concrete first step towards further AI research on this fascinating topic.

2. Related Work

Creativity and Metaphorical Abstraction. Creativity often involves an innovative fusion of objects, attributes, or relationships from previous knowledge to generate new concepts [5, 52]. Metaphors can serve as an invaluable tool for expressing creative insights and also to stimulate new ones [14, 21]. The cognitive research community has made initial attempts in understanding different realizations of metaphors such as language metaphors and visual metaphors [15, 16, 32]. Studies of visual persuasion show that visual metaphors may be more effective than language metaphors in terms of producing a greater degree of cognitive creativity [34]. In addition to improving creativity, metaphors are also known to elicit pleasure since the initial ambiguity in (re-)conceptualizing a target entity in terms of a source stimulus generates interest and motivation, and the subsequent resolution is rewarding – explaining the impor-

tance of using creative metaphorical processes in art, advertising, and marketing [6, 30, 38].

Metaphors in Language Research. Computational Linguistic studies show that metaphors are ubiquitous in language, occurring once per three sentences on average [40, 44, 49]. Recently, an increasing number of research efforts have explored the limitations and challenges in detecting and decoding the meaning of language metaphors [7, 48]. While there exist several computational models [9, 12, 17, 33, 35, 45, 47, 56] for metaphor identification, interpretation and generation in language, there exists very little work on computational modeling of visual metaphors.

Metaphors in Computer Vision. Much of computer vision literature is focused on understanding and generating literal images. Automatic metaphorical interpretations of images is highly challenging and requires multi-faceted cognitive reasoning that involves visual reasoning coupled with the use of external knowledge. Recent work on affective captioning using the ArtEmis data set [1] includes some captions that evoke metaphors to explain emotions, but that dataset is focused on visual art and does not specifically consider metaphors. There is no explicit prior work for comprehensive evaluation and development of models that can automatically comprehend or generate visual metaphors. At the same time, there exist several studies that demonstrate the potential of visual metaphors. For instance, some studies [4, 37] suggest that advertisements containing metaphorical images will be more persuasive compared to ads with language metaphors, non-metaphorical
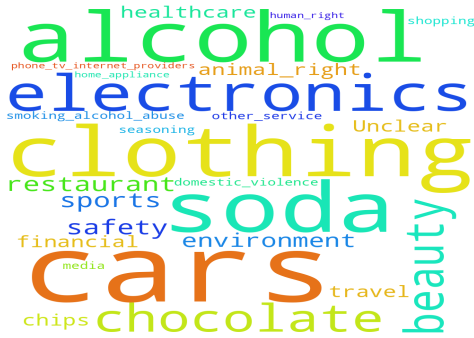
Figure 3. **Distribution of Topics** in our annotated metaphorical images from Pitt's Ad dataset.

ads, or literal images. There have been some prior computational models [11,22] for metaphor generation, but they are not thoroughly validated against any benchmark datasets or user studies. Some works [20, 54] propose datasets and techniques for general Ad image understanding with a focus on the challenging aspects of non-literal interpretations in Ad images. However they do not explicitly collect any metaphorical annotations, nor do they provide the corresponding analysis. In this work, we start with an existing Ad image dataset [20] and perform extensive human studies to filter metaphorical images and collect detailed annotations accounting various aspects of metaphoric interpretation. Prior works such as Multi-MET [57] and MET-Meme [53], propose metaphorical annotations but does not annotate the relationship between primary and secondary. In MetaCLUE, in addition to providing relationship annotations, we also collect detailed bounding box annotations that help localize the image regions invoking the primary, secondary concepts in the viewer. We further provide VQA style question and answers.

3. MetaCLUE

We introduce four different high-level tasks in Meta-CLUE that enable comprehensive evaluation and development of visual metaphor research: Classification, Localization, Understanding and gEneration. In the rest of this section, we first describe our annotation collection process for the tasks and next provide the analysis of using existing state-of-the-art techniques for each of these tasks.

3.1. Metaphor Classification

Following the fundamental vision task of image classification, we first ask: *Can we develop models that can classify whether or not a given image contains a visual metaphor?* In other words, is it possible to just use visual cues to estimate whether or not there exists a metaphorical interpretation of an image?

Annotations. For this task, we need to label whether or not each image contains a metaphor. Since we tend to see more metaphorical images in Ads, we start with images from an existing Ads dataset published by the University of Pittsburgh [20] and manually annotate whether a given Ad image contains a metaphor or not. Pitt's Ads dataset contains images of both product ads (e.g. phone ads) as well as public service announcements (e.g. forest conservation ads). Concretely, to make annotations more efficient, we use a subset of 8.5K Ad images from this dataset that are annotated to have 'symbolic' (fun, adventurous, etc.) references. We find this subset to contain a considerable portion of metaphors. Specifically, we collect 5 Yes/No annotations for each of these 8K images from different annotators. To obtain high-quality annotations, we explained what a visual metaphor is to the annotators and also conducted qualifying exams to pick final annotators for this task. We consider the images with 3 or more 'Yes' annotations as visual metaphors and the remaining as non-metaphorical. At the end, we identify 5061 images containing visual metaphors out of 8480 images. Pitt's Ads dataset [20] also comes with topic annotations (e.g., restaurant, car, animal rights etc.). Fig. 3 shows the word cloud plot of different topics in these 5061 metaphorical images. We split the metaphorical images into 3730 train and 1331 test images by maintaining the same distribution of topics in both the splits. We have two types of negative sets (non-metaphorical images) for classification experiments. One is formed by the remaining 3419 non-metaphorical images in the symbolic set, and another one is created separately by annotating an additional 3000 literal (and non-symbolic) images from Pitt's Ads dataset [20]. We add 2000 of 3419 symbolic negative images to our train split, and add the remaining 1419 images to our test split. Similarly 2000 of 3000 literal negatives are used in training, and the remaining 1000 are used for testing.

Evaluation and Results. Using our collected metaphor images, we evaluate the performance of the following state-of-the-art models in classifying an input image as metaphor: EfficientNet [46] and Vision Transformer (ViT) [13]. We fine-tune these models to classify metaphor vs. symbolic-non-metaphors and metaphor vs. literal, and use 20% of corresponding train splits for validation. Test results are reported in Table 1. Although the performance of ViT is significantly better than EfficientNet and random baselines, there is still ample room for improvement. In particular, models find it easier to distinguish metaphor images from literal images, and struggle to identify metaphors within the symbolic image pool.

3.2. Metaphor Understanding

We now describe how we collect annotations that help in capturing the metaphorical message from the images collected in previous section.

		Symbolic Neg.		Literal Neg.	
Model	#Params	Val	Test	Val	Test
Random	N/A	63.10	51.60	60.66	57.12
EfficientNet-B0	5.3M	60.76	49.67	70.94	50.30
EfficientNet-B7	66M	61.44	48.54	69.84	49.82
ViT-B/16	86M	**69.31**	**66.98**	**84.04**	**81.24**
ViT-L/16	307M	65.83	60.65	81.45	80.52

Table 1. **Accuracy of Metaphor Classification** (binary classification accuracy) using state-of-the-art classification architectures of EfficientNet [46] and ViT [13].

Annotations. We provide detailed instructions and several examples to the annotators to help them annotate primary and secondary concepts in the metaphor and also the characteristic/relationship that is transferred from secondary to the primary. We conduct multiple pilot studies to reduce the noise and to improve inter-annotator agreement. Enforcing the annotators to make sure that their annotations are linguistically readable in the following syntactic structure helped us in improving quality and consistency of the annotations: "___ is as ___ as ___", where the first blank is the primary concept, the second blank is the relationship, and the third blank is the secondary concept. Figure 2 shows some examples of these annotations. We collect 5 metaphor annotations for each image. As interpretations of metaphorical images can be highly subjective, there can exist more than one interpretation for each image, which makes it difficult to automatically remove noisy annotations. Therefore we conduct an additional human study where we show each of the annotation to five annotators and ask them to verify the correctness along three dimensions: (a) *Is the grammar correct?*; (b) *Are primary and secondary concepts correct?*; and (c) *Is the relationship correct?*. We remove annotations with a low number of votes out of 5 along each of the three dimensions, resulting in a total of over 26k clean annotations. We evaluate state-of-the-art models in understanding metaphorical message from the input images using 3 tasks namely, Retrieval, Captioning and Visual question answering.

Retrieval. The goal of this task is to retrieve the correct metaphor interpretation/statement from a candidate set given an image. In our candidate set, we choose exactly one positive (correct) metaphorical statement from its ground truth messages and uniformly sample $K - 1$ random negative statements from other images. Table 2 shows the results obtained with CLIP [39] and ALBEF [26] for K=50, reporting retrieval precision@1 and rank[4]. Although CLIP ViT-L/14 shows good zero-shot performance on random negatives with more than 76% accuracy, we observe a large drop

---

[4]Rank measures the averaged ranking value of the highest-ranked ground-truth statement with 1 being the highest possible rank.

Metaphor: Smoking cigarettes is as life-shortening as sharpening a pencil.

Captioning Result
Prediction: *Smoking is as dangerous as burning a pencil.*

Visual Question Answering Result
Q1: What is a smoker's life compared to a sharpened pencil?
Pred: *Sharp*   GT: Shortened
- - - - - - - - - - - - - - - - - - -
Q2: What is used as a visual metaphor for a sharpened pencil?
Pred: *Not smoking*   GT: Smoker's life

Metaphor: Smoking is as deadly as injecting drugs.

Captioning Result
Prediction: *Drinking and driving is as bad as injecting drugs.*

Visual Question Answering Result
Q1: What is a cigarette as compared to injecting drugs?
Pred: *Scary*   GT: Deadly
- - - - - - - - - - - - - - - - - - -
Q2: What is used as a visual metaphor for injecting drugs?
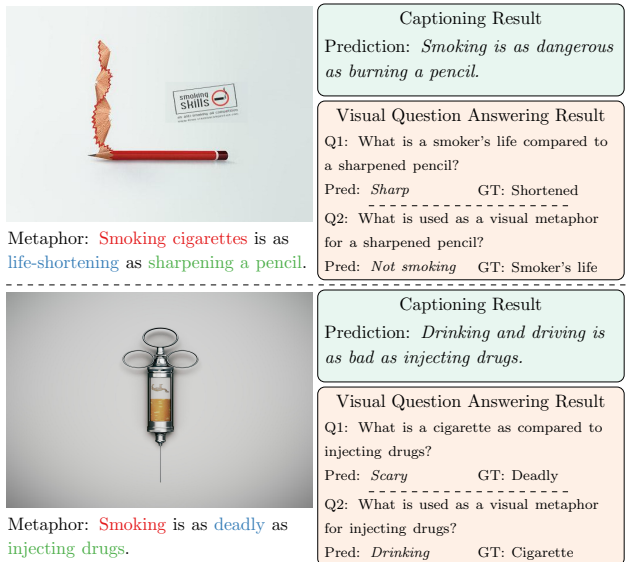Pred: *Drinking*   GT: Cigarette

Figure 4. **Results of PaLI [10]** for captioning and visual question answering on sample images in our test split.

in performance as we increase K to $\{100, 500, 1000\}$. We further fine-tune CLIP models using 70% of metaphor annotations as the train set and see gains by up to $+7$ absolute points in p@1. In summary, the performance of models is impressive with less than 50 negative candidates whereas the performance drops greatly by increasing K.

We hypothesize that models might simply be looking at salient objects rather than comprehending the underlying semantics of metaphor in finding the correct candidate. To test this, we mine hard negative (HN) statements and use them as our candidate set. Specifically, we construct the following five types of HNs: (a) **Neg Prim**: candidates obtained by replacing the primary concept in the metaphor statement with the primary concept from another image[5]; (b) **Neg Sec**: replacing secondary concept likewise; (c) **Neg Prim+Rel**: replacing primary and relationship; (d) **Neg Sec+Rel**: replacing secondary and relationship; (e) **Swap Prim&Neg**: swapping primary and secondary from the same image[6]. Table 2 shows the results. We see a significant drop of up to 30% by using HNs as negative statements for $K = 50$, indicating the difficulty in comprehending and distinguishing metaphorical abstraction of concepts. We find performance of **Neg Prim** is significantly higher than **Neg Sec**, suggesting that the models tend to rely more on primary object than secondary object in identifying correct interpretation. Overall, there is ample room for improvement indicated by the steep drop in model performance with HNs.

Captioning. Here, we propose metaphor image cap-

---

[5]We swap objects with images having different topic to make sure that the generated HNs are actually negatives.

[6]With swapping K is always 2.

Model	Random Neg p@1 ↑	Random Neg rank ↓	Neg Prim p@1 ↑	Neg Prim rank ↓	Neg Sec p@1 ↑	Neg Sec rank ↓	Neg Prim+Rel p@1 ↑	Neg Prim+Rel rank ↓	Neg Sec+Rel p@1 ↑	Neg Sec+Rel rank ↓	Swap accuracy ↑
CLIP (ViT-B/16)	70.97	3.49	46.67	10.11	38.14	13.97	49.36	8.80	42.25	11.69	40.61
CLIP (ViT-B/32)	61.78	4.19	38.74	11.60	33.20	14.84	39.79	10.29	36.27	12.71	41.28
CLIP (ViT-L/14)	76.66	3.17	51.75	9.22	39.86	13.19	54.74	7.70	45.99	11.02	43.08
ALBEF	39.79	8.57	27.00	16.53	29.31	15.79	26.77	15.58	28.42	14.59	46.67
ALBEF (MSCOCO)	44.87	7.55	31.78	15.13	31.41	14.09	32.53	14.04	33.28	12.85	48.24
ALBEF (Flickr30k)	47.49	8.77	35.60	14.56	35.22	13.66	36.35	13.58	36.12	12.79	**49.81**
FT CLIP (ViT-B/16)	76.81	2.25	49.81	9.65	45.47	10.24	53.40	8.07	50.63	8.18	44.65
FT CLIP (ViT-B/32)	68.06	2.82	43.00	10.82	39.64	11.23	44.72	9.39	43.45	9.35	45.69
FT CLIP (ViT-L/14)	**81.75**	**1.78**	**57.66**	**7.48**	**49.06**	**9.40**	**61.25**	**5.99**	**57.06**	**7.33**	43.75

Table 2. **Performance of retrieval models** on K random (column 2) and hard negative candidates (columns 3-7) ($K = 50$ for columns 2-6, $K = 2$ for last column).

Captioning	Acc	BLEU4	ROUGE-L	METEOR	SPICE	CIDEr
Whole caption	1.1%	0.254	0.536	0.220	0.186	1.076
Primary	29.9%	0.327	0.407	0.338	0.244	0.931
Secondary	13.7%	0.249	0.307	0.193	0.155	0.550
Relationship	23.6%	0.485	0.203	0.226	0.028	0.296

VQA	Acc	BLEU4	ROUGE-L	METEOR	SPICE	CIDEr
All questions	19.9%	0.329	0.286	0.249	0.185	0.851
Primary	21.5%	0.291	0.348	0.277	0.290	1.099
Secondary	12.8%	0.238	0.232	0.181	0.234	0.735
Relationship	25.6%	0.449	0.275	0.285	0.038	0.706

Table 3. **Metaphorical image captioning (left) and visual question answering (right) performance of PaLI** [10]. We report different metrics (higher the better) using both the exact match accuracy (Acc) and standard text generation metrics. For image captioning, we evaluate the whole predicted caption as well as parsed primary object, secondary object, and relationship. For VQA, we evaluate on all predicted answers as well as provide the breakdown for each question type.

tioning task, where the input is an image and the target is the metaphorical message in the syntactic structure `<primary_concept> is as <relationship> as <secondary_concept>`. We fine-tune and evaluate the state-of-the-art literal image caption model PaLI-17B [10] based on the exact match accuracy (maximum over all references) and standard metrics for image captioning BLEU4 [36], ROUGE-L [28], METEOR [3], SPICE [2], and CIDEr [51]. Since the target captions follow the fixed syntactic structure, we parse each predicted caption into the primary concept, the secondary concept, and their relationship, and use the same set of metrics for the whole caption to evaluate.[7] Table 3 (left) summarizes the results and Figure 4 provides sample qualitative results. We observe that PaLI generally struggles on this task. For instance, it achieves a CIDEr score of 1.076, compared to 1.491 for state-of-the-art literal image captioning on the popular COCO-Captions [8]. Further, the model struggles with predicting the target relationship when the metrics favor recall (e.g., ROUGE-L) and with predicting the target secondary concept when the metrics favor precision (e.g., BLEU).

Visual Question-Answering (VQA). We propose metaphorical open-ended (i.e., not vocab-based) VQA task, where the input is an image and a given question, and

the target is the answer. We use fixed templates to generate 2 VQA questions whose answer is the primary concept, 2 for secondary concept, and 2 for relationship. Again, we fine-tune and evaluate the state-of-the-art literal VQA model PaLI-17B [10], using the same set of metrics as in image captioning. Table 3 (right) summarizes the results and Figure 4 provides sample qualitative results. Overall, we find that PaLI performs poorly, only achieving the average accuracy score of 19.9%, while the state-of-the-art literal VQA on the popular VQAv2 [18] benchmark is 77.6% on "other" questions. Additionally, the model struggles the most with answering questions that ask for the secondary concept.

3.3. Metaphor Localization

Here, the task is to localize the image regions that invoke either the primary or secondary concept in the viewer. This is similar to the phrase grounding task of localizing objects using free-form natural language phrases [23, 31, 55], but with some key differences due to the peculiarities of visual metaphors in comparison to literal images used in standard vision datasets.

Annotations. As discussed earlier, there are diverse types of metaphors based on how the primary and secondary concepts are visually depicted in an image (see Fig. 2). There are at least two key differences in metaphor localization

---

[7]The score of 0 is given if parsing fails.

The shoe is as light as feathers.

Ketchup is as hot as a volcano.

Getting two ice creams is as exciting as getting two movie tickets for one.
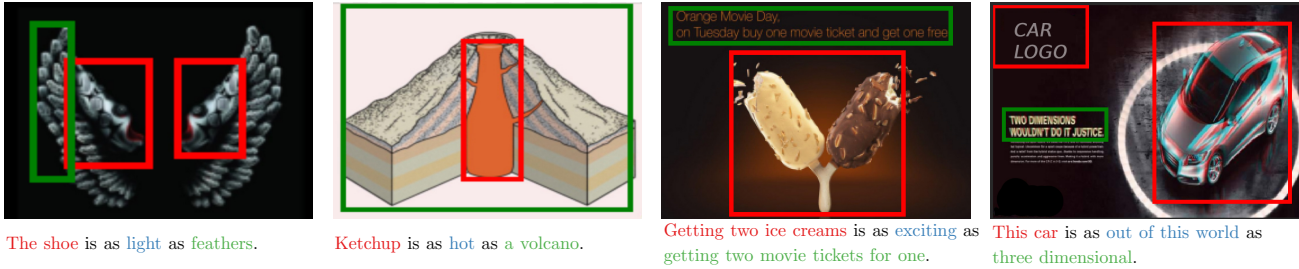
This car is as out of this world as three dimensional.

Figure 5. **Sample Localization annotations** showing annotated bounding boxes around primary and secondary concepts. Notice the diversity of types in bounding boxes: explicit, contextual, logo and texts. This makes metaphor localization more challenging compared to standard object detection.

Parking assist is as safe as parking far from objects.

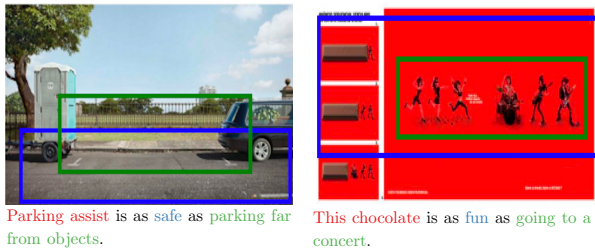This chocolate is as fun as going to a concert.

Figure 6. **Sample Localizations** with the phrase-grounding model from [27], where the secondary concepts are contextual. GT boxes are shown in green, whereas the predictions are shown in blue.

	mAP_{50}	mAP_{70}
Primary concept	33.22	14.25
Secondary concept	43.54	31.23

Table 4. **Localization results** with CLIP based phrase localization model [27] on our test split.

compared to standard localization in literal images: 1. A given concept can be present in the image either explicitly or in a contextual manner (for e.g., contextual visual metaphors in Fig. 2). 2. Visual metaphor Ads are inherently multimodal and a concept can be invoked by other modalities such as text or logo. See the multimodal visual metaphors in Fig. 2. As a result, we not only annotate the bounding boxes that invoke the primary/secondary concept in the viewer, but we also annotate the type of that bounding box. A bounding box can be of one of the 4 types: Explicitly present, Contextually present, Logo or a Text. Specifically, for each of the 5061 metaphorical images, we pick the best metaphor annotation (primary, secondary concepts and their relationship) according to their validation scores (see previous section) and collect bounding box annotations for both the primary and secondary concepts. We collect all the bounding boxes that invoke both the primary and secondary concepts and also their type (explicit, contextual, logo or text) for each of the images. We use 5 annotators for each annotation and choose the bounding boxes with the best inter-annotator agreement. We use the same train and test splits as used in understanding tasks. Fig. 5 shows sample localization annotations. We collected over 30k bounding box annotations for this task on 5061 metaphor images.

Detection. Recent detection and localization models [27, 58] pre-trained on image and caption pairs are shown to

achieve remarkable localization performance on discriminating fine-grained objects and unseen concepts. Specifically, we evaluate [27] which leverages the effective image representations in CLIP by extracting spatial features from it. Using these spatial features, for each pixel location, the model computes the inner product between the spatial feature and the phrase embedding extracted from CLIP to predict the bounding box. In our case, we pass the primary or secondary concept as input phrase to [27] the estimate the corresponding bounding box. Table 4 summarizes the detection results (using Mean Average Precision) on our test split. We find relatively better performance in localizing secondary objects compared to primary objects. We show few qualitative results in Figure 6. It is worth noting that our collected annotations allow for more comprehensive analysis on localization tasks due to the availability of different types of bounding boxes (explicit, contextual, logos and texts).

3.4. Metaphor Generation

Recent large-scale text-to-image (T2I) generative models show remarkable success in generating highly realistic images from text prompts. *Can these models also work well in metaphorical image generation?* We evaluate two state-of-the-art generative models (Imagen [42], Stable diffusion [41] (SD)) using 300 samples from MetaCLUE test set where we use the text prompts: "An advertisement where *primary-concept* is as *relationship* as *secondary concept*." In addition, we finetune the stable diffusion model on our train split (same split as in Sec. 3.2).

Results. Fig. 7 shows sample visual results from different T2I models, with the metaphor annotation shown on the

Model	FID ↓	CLIP Similarity ↑
Imagen [42]	153.1	32.1
Stable Diffusion [41]	161.6	30.8
Stable Diffusion - FT	154.3	32.0

Table 5. **Analysis of Image Generation** results with standard metrics of FID [42] and CLIP similarity [39] scores.

top. The generated images capture different aspects of the metaphor (tablet, waterproof), but not the entire metaphorical message. We compute standard metrics to automatically evaluate the quality of the generations. Tab. 5 shows the standard FID [19] and CLIP-Similarity [39] scores of different models. FID score evaluates the image distribution similarity between the generation images with the corresponding real image distribution. FID scores in Tab. 5 shows that Imagen performs slightly better than SD. And, there is a slight improvement in FID with finetuning (SD-FT vs. SD). In general, high FID scores in Tab. 5 indicate the large distribution gap between the generated and real images. CLIP similarity score, on the other hand, measures the prompt fidelity - similarity between the generated image and the corresponding input text prompt according to the CLIP model [39]. CLIP similarity scores in Tab. 5 follow the similar trend as FID scores: Imagen performs better than SD in terms of the prompt fidelity, and finetuning SD model improved its prompt fidelity.

Given that both FID and CLIP scores are not tailored towards metaphorical images, these metrics are not reliable in assessing whether the generated images capture the essence of visual metaphors. To analyze this, we perform human studies comparing two different models at a time. Specifically, we show a metaphorical message and the corresponding generations from two models and ask the users to pick an image that best conveys that metaphorical message. We obtain 7 user ratings for each image pair and consider 3-out-of-7 or 4-out-of-7 to be ties. Fig. 8 shows the percentage of user preferences across different pairs of results. User studies also indicate that Imagen performed better than SD. An interesting finding is that finetuning SD resulted in slightly worse user preference compared to base SD model (SD-FT vs. SD). This is in contrast to standard FID and CLIP metrics that improved with finetuning. We hypothesize that finetuning SD resulted in more realistic Ad images, but the resulting model may have forgotten some of the metaphorical priors. This calls for more effective finetuning strategies with the limited training datasets, which forms an important future work. In addition, both Imagen and SD performed quite poorly compared to real images in conveying metaphorical messages. Real images are preferred around 88% of time over Imagen results. This illustrates the big scope of improvements in generating visual metaphors.

Metaphor: This android tablet is as waterproof as someone in a swimtube.

Real Imagen

Stable Diffusion Stable Diffusion - FT

Figure 7. **Sample Image Generations** for a given metaphorical message (shown on top) with Imagen [42], Stable Diffusion [41] and fine-tuned (FT) version of Stable Diffusion.
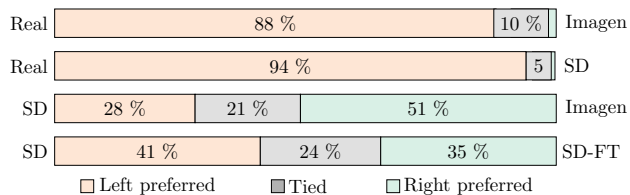

Figure 8. **User Study on Image Generation Results.** Percentage of results users preferred across real, Imagen [42], Stable Diffusion (SD) [41] and its fine-tuned version (SD-FT) results. Users are asked to choose the image that better depicts a given metaphor.

4. Conclusion

In this paper, we present a step towards comprehensive evaluation of progress on visual metaphor research. Specifically, we propose a collection of tasks related to comprehending and generating visual metaphors using AI techniques. Our MetaCLUE tasks include Classification, Understanding (Retrieval, Captioning, VQA), Localization and Generation. For comprehensive evaluations, we collected high quality and rich annotations that facilitate the measurable progress. Existing methods demonstrate poor results in many cases with our experimental analysis shedding light on strengths and drawbacks of different approaches paving a path for future research in this fascinating field.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. 2, 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *ECCV*, 2016. 6

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005. 6

[4] Roland Barthes. *Image-music-text*. Macmillan, 1977. 3

[5] Nathalie Bonnardel and Evelyne Marmèche. Towards supporting evocation processes in creative design: A cognitive approach. *International journal of human-computer studies*, 63(4-5):422–435, 2005. 3

[6] Robert W Boozer, David C Wyld, and James Grant. Using metaphor to create more effective sales messages. *Journal of services marketing*, 1990. 3

[7] Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, 2021. 2, 3

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[9] Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: Ets team in 2020 metaphor shared task. In *Proceedings of the second workshop on figurative language processing*, pages 235–243, 2020. 3

[10] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 5, 6

[11] Lydia B Chilton, Savvas Petridis, and Maneesh Agrawala. Visiblends: A flexible workflow for visual blends. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. 4

[12] Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, 2021. 2, 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5

[14] Gilles Fauconnier and Mark Turner. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books, 2008. 3

[15] Charles Forceville. *Pictorial metaphor in advertising*. Routledge, 1996. 2, 3

[16] Susan R Fussell and Mallie M Moss. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141, 1998. 3

[17] Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153, 2020. 3

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 6

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8

[20] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 2, 3, 4

[21] Bipin Indurkhya. On the role of metaphor in creative cognition. In *ICCC*, pages 51–59, 2010. 3

[22] Bipin Indurkhya and Amitash Ojha. An empirical study on the role of perceptual similarity in visual metaphors and creativity. *Metaphor and Symbol*, 28(4):233–253, 2013. 4

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[24] George Lakoff. The contemporary theory of metaphor. 1993. 2

[25] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008. 2

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 5

[27] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022. 7

[28] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 6

[29] Todd I Lubart and Isaac Getz. Emotion, metaphor, and the creative process. *Creativity research journal*, 10(4):285–301, 1997. 2

[30] Carita Lundmark. *Metaphor and creativity in British magazine advertising*. PhD thesis, Luleå tekniska universitet, 2005. 3

[31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6

[32] James H Martin. A corpus-based analysis of context effects on metaphor comprehension. *Trends in Linguistics Studies and Monographs*, 171:214, 2006. 3

[33] Zachary J Mason. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44, 2004. 3

[34] Edward F McQuarrie and David Glen Mick. Visual rhetoric in advertising: Text-interpretive, experimental, and reader-response analyses. *Journal of consumer research*, 26(1):37–54, 1999. 3

[35] Ekaterina Ovchinnikova, Vladimir Zaytsev, Suzanne Wertheim, and Ross Israel. Generating conceptual metaphors from proposition stores. *arXiv preprint arXiv:1409.7619*, 2014. 3

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. 2002. 6

[37] Barbara J Phillips. The impact of verbal anchoring on consumer response to image ads. *Journal of advertising*, 29(1):15–24, 2000. 3

[38] Barbara J Phillips and Edward F McQuarrie. Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing theory*, 4(1-2):113–136, 2004. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5, 8

[40] Sunny Rai and Shampa Chakraverty. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37, 2020. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 7, 8

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 7, 8

[43] Linda M Scott. Images in advertising: The need for a theory of visual rhetoric. *Journal of consumer research*, 21(2):252–273, 1994. 2

[44] Gerard Steen. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing, 2010. 3

[45] Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 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, 2021. 2, 3

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4, 5

[47] Asuka Terai and Masanori Nakagawa. A computational system of metaphor generation with evaluation mechanism. In *International Conference on Artificial Neural Networks*, pages 142–147. Springer, 2010. 3

[48] Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, 2021. 2, 3

[49] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, 2011. 3

[50] Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160, 2016. 2

[51] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[52] Merryl J Wilkenfeld and Thomas B Ward. Similarity and emergence in conceptual combination. *Journal of Memory and Language*, 45(1):21–38, 2001. 3

[53] Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2887–2899, 2022. 4

[54] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 837–855, 2018. 4

[55] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 6

[56] Zhiwei Yu and Xiaojun Wan. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 861–871, 2019. 3

[57] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 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, 2021. 4

[58] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 7