

# Supplementary Material for 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

In this supplementary material, we provide additional details on our data annotation process and experiments.

## A. Metaphor Classification

In collecting annotation labels for whether or not each image contains a metaphor, we use 5 annotators per image. All annotators are based in the United States. To increase quality of the annotations, we provide detailed instructions to the annotators on visual metaphors and also conduct qualifying exams to pick final annotators for this task. Specifically, we show multiple examples to help annotators clearly understand the notion of literal and metaphor images; primary, secondary objects and the metaphorical relationships. Figure 2 shows an example of instructions that we show to annotators. We additionally conduct multiple pilot studies to identify the most effective method of collecting the annotations. In Figure 3, we show our final template used in collecting the annotations. The placeholder image in the template will be replaced by Ad image in the study. We find providing the ground-truth messages (i.e. the Image Ad description provided in the Pitt’s Ads dataset images [1]) to the annotators further helps in improving the quality and consistency of annotations. We consider the images with 3 or more ‘Yes’ annotations as visual metaphors and the remaining as non-metaphorical. Figure 1 shows few qualitative examples for the predictions obtained by ViT-L/16.

**Is image-only classification possible with metaphorical images?** We performed a pilot study on a random sample of 300 images (containing symbolic and metaphor images) in manually annotating whether the image is a metaphor with out providing ground-truth description and we find 89.6% of samples could be correctly classified as metaphors. The percentage becomes 92.3% when it comes to classifying metaphors from literal images - indicating that image-only classification is feasible. The ability of the existing models to jointly process image and text to comprehend metaphors is tested in our retrieval and VQA tasks to some extent.

\*Equal Contribution

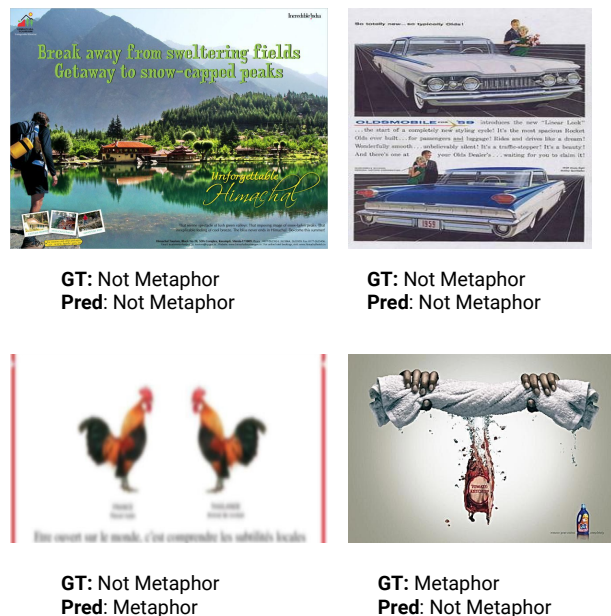


Figure 1. **Qualitative results** for ViT-L/16 classification model (we intentionally blurred the bottom-left image).

## B. Metaphor Understanding

After we collected the metaphor images, we conduct next phase, Phase 2, of study where we ask annotators to describe the metaphor in a sentence of the form “(primary concept) is as (relationship) as (secondary concept)”. Annotators are again required to pass a qualifying exam in order to participate in this phase. In this exam, annotators are asked to select the correct primary concept, secondary concept, and relationship for few metaphorical images. About 50 annotators passed this exam. In Figure 4, we show our final template used in collecting the annotations.

We also conduct an additional validation phase, Phase 3, where we gave human raters a metaphorical image and a corresponding sentence collected in Phase 2, and ask them

to validate whether the sentence is a correct annotation of the metaphor present in the image. Correctness is evaluated across 3 dimensions: correct grammar, correct primary/secondary concept, and correct relationship. Figure 5 shows our template used in the validation phase. Annotators are again required to pass an exam in order to qualify for this phase, although this exam is not as difficult as the exam in Phase 2. About 400 Annotators passed this exam. We used 5 Annotators per (image, annotation) pair from Phase 2.

### B.1. Retrieval

In our candidate set for retrieval task, we choose one positive (correct) metaphorical statement from its ground truth messages and uniformly sample  $K - 1$  random negative statements from other images. Figure 6 compares the performance of models with several different sizes  $K$  of the candidate set ranging from  $K = 10$  to 1000. The performance of models is impressive with less than 50 negative candidates whereas the performance drops greatly by increasing  $K$ . Table 2 in the main paper show the results with hard negatives for  $K = 50$ . We compare the performance on hard negatives by changing  $K$  in Figure 7, Figure 8, Figure 9, and Figure 10. The human baseline for our retrieval task (on a sample of 200 randomly selected images) is 97.5%.

### B.2. VQA

We use the following template for creating VQA datasets from the caption of the form `<primary> is as <relationship> as <secondary>`. We generate 2 questions per answer type target. The human baseline for our VQA task (on a sample of 200 randomly selected images) is 94.0%.

- What is used as a visual metaphor for `<secondary>`? `<primary>`
- What is as `<relationship>` as `<secondary>`? `<primary>`
- What is `<primary>` a visual metaphor for? `<secondary>`
- What is `<primary>` as `<relationship>` as? `<secondary>`
- What is `<primary>` as compared to `<secondary>`? `<relationship>`
- What is `<primary>` like compared to `<secondary>`? `<relationship>`

### C. Localization

We pick the best metaphor annotations (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 allow more than 1 box to be drawn for each of the 4 types. For the primary and secondary concept, annotators are asked to mark whether the concept is “explicitly visible” in the image, or only present “contextually”. We use 5 annotators per example and conduct pilot studies to identify the best possible way to collect annotations. Figure 11 shows our template used in collecting the localization annotations. We perform a filtering on these collected localization annotations to remove overlapping bounding boxes for each of the 4 types of boxes. For each image, we determine which bounding boxes have an Intersection-over-Union (IoU) of greater than 0.5. The distinct boxes with the most overlaps are preserved; all other boxes are discarded. Using this approach, it is possible for a given image and box type to have more than one bounding box. In computing Mean Average Precision of localization models on localizing Explicit and Contextual regions, we pick the max IoU among all ground-truth boxes as the final score.

It may be noted that we did not collect annotations to differentiate between contextual, hybrid, multimodal, juxtaposition categories, which we find to be difficult to annotate with high inter-annotator agreement. Instead, we collected annotations on whether or not the primary, secondary concepts are explicit or contextually implied. We found at least 4 out of 5 annotators agree on 76.1% explicit boxes and 40.6% contextual boxes.

### D. Generation

Figure 12 and Figure 13 show additional qualitative results from different text-to-image generation models. It may be noted that providing a literal description of metaphor could help in generating better images. Our experiments also make this point that current large generative AI models cannot directly interpret metaphorical messages. Automatically converting a metaphorical message into a visual description is non-trivial, which we leave to future work.

### References

- [1] 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] 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. 9

- [3] 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. 9

### What is a Visual Metaphor?

A "visual metaphor" is an image which shows 2 objects or actions which have some relationship with each other.

Of these 2 objects / actions, one of them is called the "primary concept", and the other is called the "secondary concept".

The "primary concept" is the main focus of the image.

The "secondary concept" has some relationship with the primary concept.

Note that both the "primary" and "secondary" concept can be either an OBJECT (a noun) or an ACTION (a verb).

Furthermore, in order for an image to be a visual metaphor, you MUST be able to construct a coherent sentence using the following format: "\_\_\_\_\_ is as \_\_\_\_\_ as \_\_\_\_\_." The 1st blank is the primary concept, the 2nd blank is the relationship, and the 3rd blank is the secondary concept.

To understand this better, let's look at some examples.

#### Examples of visual metaphors:

Example #1:

Primary: "phone"  
Secondary: "rocket ship"  
Relationship between them: "fast"  
As a sentence: "THE PHONE is as FAST as A ROCKET SHIP."



Example #2:

Primary: "shoes"  
Secondary: "fitting an unusual foot"  
Relationship between them: "customizable"  
As a sentence: "THE SHOE is as CUSTOMIZABLE as FITTING AN UNUSUAL FOOT."



#### Literal secondary concept vs Symbolic secondary concept

In the above examples, both the primary concept and the secondary concept in the image.

However, in some images, the secondary concept will not directly appear in the image, but will only be referred to symbolically, or by text written on the image.

The secondary concept does NOT have to be present in the image. Sometimes, an image will have 2 or more concepts present in it, but none of those concepts will be the correct secondary concept.

For example, in the following image, the 2 concepts that are present are "car" and "glass". "car" is the correct primary concept. However, the CORRECT secondary concept is actually "beetle", because the car trapped under the glass symbolizes a beetle trapped under a glass. In other words, the relationship between "car" and "beetle" is "they are trapped under a glass".



#### More examples in which the secondary concept is not directly visible:

Example #1:

Primary: "hockey player wearing pads"  
Secondary: "seat belt" (notice that the seat belt doesn't appear in the image; it is only referenced by the text "buckle up")  
Relationship between them: "safe"  
As a sentence: "A HOCKEY PLAYER WEARING PADS is as SAFE as A SEAT BELT."



Example #2:

Primary: "elephant"  
Secondary: "wall with graffiti on it" (notice that the wall doesn't appear in the image)  
Relationship between them: "prone to damage"  
As a sentence: "AN ELEPHANT is as PRONE TO DAMAGE as A WALL WITH GRAFFITI ON IT."




#### Example of an image that does NOT contain a visual metaphor:

This example does NOT contain a visual metaphor (there is only 1 concept, the car):



Figure 2. Detailed annotation instructions used in our human study to help annotators familiarize with the metaphor concepts.

Now, consider the following image advertisement:



Additionally, here are some descriptions of the image:

\*Groundtruth 1"

\*Groundtruth 2"

- Creative is Missing
- Creative Only Partially Loaded
- Wrong language
- Unexpected Porn


\* Does the image contain a visual metaphor?

Yes

No

Figure 3. **Human-study template** used for filtering metaphorical images.

Now, consider the following image advertisement:



Additionally, here are some descriptions of the image:

\*Groundtruth 1"

\*Groundtruth 2"

- Creative is Missing
- Creative Only Partially Loaded
- Wrong language
- Unexpected Porn

\* Is the secondary concept visually present in the image? Or is it only referred to symbolically in the image?

Visually present

Only referred to symbolically

\* Now, please fill in the blanks in the following sentence: "\_\_\_\_\_ is as \_\_\_\_\_ as \_\_\_\_\_." The 1st blank is the primary concept, the 2nd blank is the relationship, and the 3rd blank is the secondary concept.

is asas

Figure 4. **Phase 2 data annotation template** for collecting primary, secondary and relationship annotations.

Now, consider the following image advertisement:



Additionally, here are some descriptions of the image:

"Groundtruth 1"

"Groundtruth 2"

- Creative is Missing
- Creative Only Partially Loaded
- Wrong language
- Unexpected Porn

Now, consider the following sentence:  
"Annotation"

\* Is the grammar correct in this annotation?

Yes

No

\* Are the primary and secondary concepts correct in this annotation?

Yes

No

\* Is the relationship between primary and secondary concepts correct in this annotation?

Yes

No

Figure 5. **Phase 3 data annotation template** for validating the annotations.

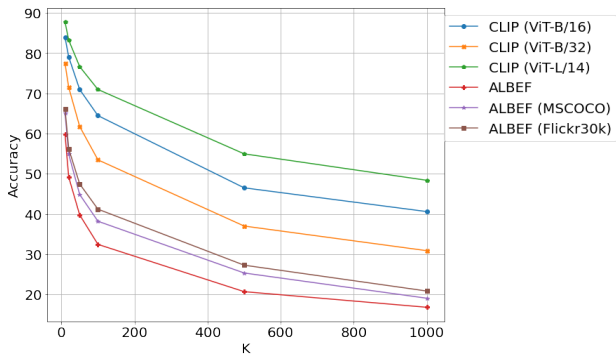


Figure 6. Performance of retrieval models on  $K$  random negative candidates.

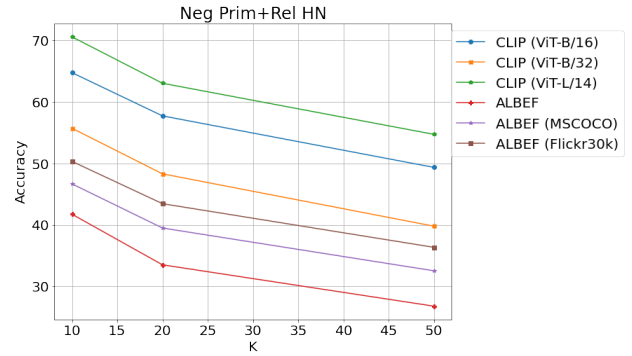


Figure 9. Performance of retrieval models on  $K$  Neg Prim+Rel hard negative candidates.

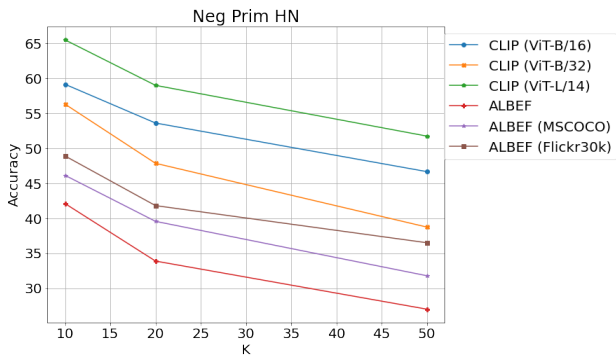


Figure 7. Performance of retrieval models on  $K$  Neg Prim hard negative candidates.

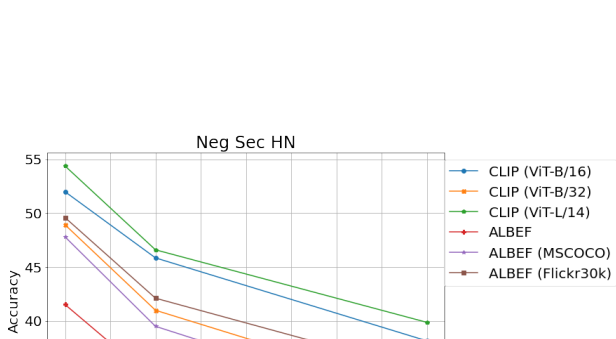


Figure 8. Performance of retrieval models on  $K$  Neg Sec hard negative candidates.

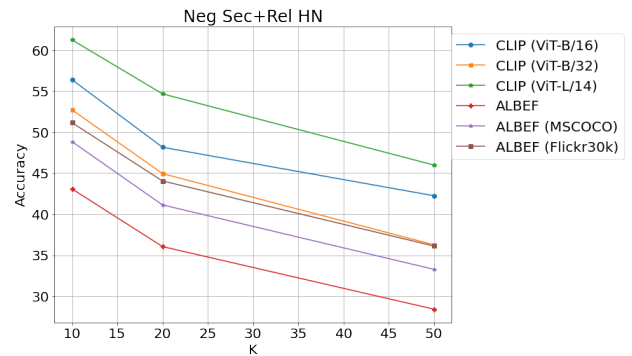


Figure 10. Performance of retrieval models on  $K$  Neg Sec+Rel hard negative candidates.

Now, consider the following image advertisement:



Additionally, here are some descriptions of the image:

\*Groundtruth 1\*

\*Groundtruth 2\*

- Creative is Missing
- Creative Only Partially Loaded
- Wrong language
- Unexpected Porn

Here is a sentence which describes the visual metaphor present in this image:

**"The car is as adventurous as a spaceship"**

This is the PRIMARY concept:

**"The car"**

This is the SECONDARY concept:

**"a spaceship"**

Is the PRIMARY concept explicitly visible in the image? Or, is it not explicitly visible but implied by the image's context?

- Explicitly visible
- Not explicitly visible

\* Draw a box around the PRIMARY concept.



Is the SECONDARY concept explicitly visible in the image? Or, is it not explicitly visible but implied by the image's context?

- Explicitly visible
- Not explicitly visible

\* Draw one or more boxes around all parts of the image which are contextually relevant to the SECONDARY concept.



Does this image contain a logo? Is this logo relevant to the metaphor?

- There is a logo, and it is relevant to the PRIMARY concept
- There is a logo, and it is relevant to the SECONDARY concept
- There is a logo, and it is relevant to BOTH the PRIMARY and SECONDARY concept
- There is a logo, but it is not relevant
- There is no logo

\* Draw a box around the logo.



Does this image contain overlaid text? Is this text relevant to the metaphor?

- The image contains text, and it is relevant to the PRIMARY concept
- The image contains text, and it is relevant to the SECONDARY concept
- The image contains text, and it is relevant to BOTH the PRIMARY and SECONDARY concept
- The image contains text, but it is not relevant
- The image does not contain text

Figure 11. **Human-study template** used for annotating bounding boxes of metaphor concepts.



Metaphor: This vehicle is as capable of traversing as ants.



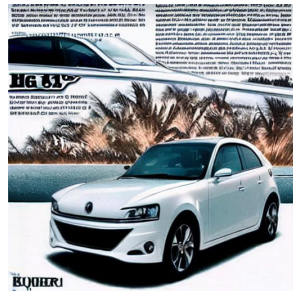
Real



Imagen



Stable Diffusion



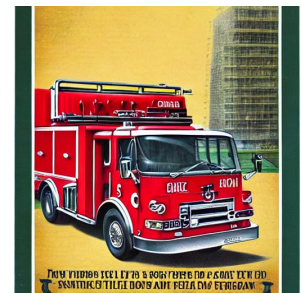
Stable Diffusion - FT

Figure 12. More Qualitative Results for Image Generations for a given metaphorical message (shown on top) with Imagen [3], Stable Diffusion [2] and fine-tuned (FT) version of Stable Diffusion.

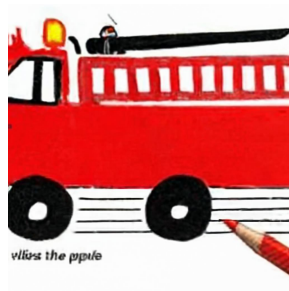
Metaphor: This pencil is as red as fire truck.



Real



Imagen



Stable Diffusion



Stable Diffusion - FT

Figure 13. More Qualitative Results for Image Generations for a given metaphorical message (shown on top) with Imagen [3], Stable Diffusion [2] and fine-tuned (FT) version of Stable Diffusion.