

## A APPENDIX

### A.1 DATASET ACCESS

ObjectNet Captions is contained in the supplemental material and will be available for download publicly upon acceptance. ObjectNet Captions inherits the ObjectNet license which is derived from Creative Commons Attribution 4.0. Importantly, the license prohibits use of the dataset for updating the parameters of a model. The authors bear all responsibility in case of violation of rights, etc., and confirmation of the data license. The ObjectNet webpage is frequently maintained and updated. Backup Dropbox download links are also provided.

### A.2 COMPUTING HUMANr

We have published a toolkit for effortlessly launching caption comparison tasks on Amazon Mechanical Turk for computing HUMANr. HUMANr can be run on any dataset and model with only a handful of command-line instructions. The code is released in the supplemental material and will be made public upon acceptance.

### A.3 DATASET EXAMPLES

In appendix [A.3](#) we show some more examples from ObjectNet Captions along with human caption comparison judgments.

### A.4 MODEL SPECIFICATIONS

All models weights are publicly available.  $GIT_L$  model weights can be found at <https://github.com/microsoft/GenerativeImage2Text>. We use the `GIT_LARGE_COCO` variant. ExpansionNetv2 weights can be found at [https://github.com/jchenghu/expansionnet\\_v2](https://github.com/jchenghu/expansionnet_v2), and ClipCap weights can be found at [https://github.com/rmokady/clip\\_prefix\\_caption](https://github.com/rmokady/clip_prefix_caption). We conducted no parameter tuning of any model in any of our experiments. Model inference was performed on a cluster of 8 Nvidia TITAN RTX graphics cards.

### A.5 DATASET ANALYSIS

The following figures show additional analysis of the linguistic properties of ObjectNet captions including longer captions with linguistic diversity and complexity compared to other captioning datasets.

### A.6 EVALUATION RESULTS

### A.7 WORKER COMPENSATION

For each task, rewards were chosen to estimate a \$15/hr wage. In practice, average wage for workers on each task was well above this. An estimated total of \$15,000 was spent collecting ObjectNet Captions and another \$3,000 was spent to collect HUMANr results for humans and our 3 models on 3 datasets. We note that stable and reproducible HUMANr results can be collected with far fewer comparisons, reducing cost by up to 10 times compared to our experiments.

### A.8 EXPERIMENTAL INSTRUCTION

Workers were consented for all tasks and were given clear instruction. Screenshots of consent and instruction pages for transcription and HUMANr tasks are attached below.

|   |                           |   |  |              |
|---|---------------------------|---|--|--------------|
|    | Human 1 vs Human 2        | A pair of black pants are laying across the ground carpet. I see a couple of little kids' balls sitting there.  | A pair of black pants laid out on a brown rug.   | HUMANr score |
|   | Human vs GIT <sub>L</sub> | Black yoga pants laying on a brown carpet and what appears to be the living room with two small balls next to it.   | this image may contain clothing apparel pants denim and jeans  | -1.00        |
|   | Human vs ExpansionNetv2   | A pair of black pants is sitting on top of a brown carpeted floor. Also on the floor is a small basketball and a small soccer ball. I can see a white wall with a white baseboard and the background. | A person s feet on the floor with a soccer ball.   | 0.00         |
|   | Human vs ClipCap          | A black pair of pants laying on a brown.  | A pair of black leather pants sitting on top of a floor.   | 0.00         |
|   | Human 1 vs Human 2        | There's a red and black item laying on top of a white sink in a bathroom.   | A black and red Bluetooth speaker sitting on the white safe.   | HUMANr score |
|   | Human vs GIT <sub>L</sub> | A travel-size toothbrush bag on a white sink top. Also a tube of toothpaste can be seen in the top of the picture and a black bowl.   | the case is made of plastic.   | -1.00        |
|   | Human vs ExpansionNetv2   | A travel-size toothbrush bag on a white sink top. Also a tube of toothpaste can be seen in the top of the picture and a black bowl.   | A black and white blood on a bathroom sink.  | -0.75        |
|   | Human vs ClipCap          | A black and red Bluetooth speaker sitting on the white safe.  | A red object sitting on top of a wooden table.   | -1.00        |
|  | Human 1 vs Human 2        | Someone took a picture of their book the book is titled Missing 411 it's a red book on it's on the floor and the floor is dark.   | A paperback book titled Missing 411 is sitting on top of a brown tiled floor.  | HUMANr score |
|   | Human vs GIT <sub>L</sub> | There's a red and blue book titled The Missing 411 sitting against a beige background its crumpled umm folded up at the bottom corner.  | this is a book lying on the floor.   | -0.75        |
|   | Human vs ExpansionNetv2   | A paperback book titled Missing 411 is sitting on top of a brown tiled floor.   | A book sitting on the floor.   | -1.00        |
|   | Human vs ClipCap          | A paperback book, entitled "Missing 411," is sitting on top of a brown carpeted floor.  | A book with a picture of a person holding a book.  | -1.00        |
|  | Human 1 vs Human 2        | There is a used spatula being held by a man.  | A person is holding a white plastic spatula in their hand. The spatula is held over a white tile kitchen floor. I can see the barefoot of the person holding the spatula. I also see a rug on the floor. | HUMANr score |
|   | Human vs GIT <sub>L</sub> | Somebody holding up a white spatula above a brown tile floor.   | a person holding a spatula in their hand.  | -0.50        |
|   | Human vs ExpansionNetv2   | A dirty spatula with a guy's hand looks like his scars all over his arm and a dirty background floor with a half of a rug in it.  | A person holding a white spoon on the floor.   | 0.50         |
|   | Human vs ClipCap          | There is a used spatula being held by a man.  | A person holding a white frisbee in their hand.  | 0.50         |
|   |                           |   |  | -1.00        |

Figure 5: The first column contains human-generated captions and the second column contains captions from either a model or a second human. The HUMANr score in the third column indicates which caption best matched the image, as determined by another human rater. A score of -1 favors the first caption, a score of 1 favors the second caption, and a score of 0 indicates a tie. Green highlights the preferred captions, and grey indicates the less preferred caption. Both captions are green if equally preferred.

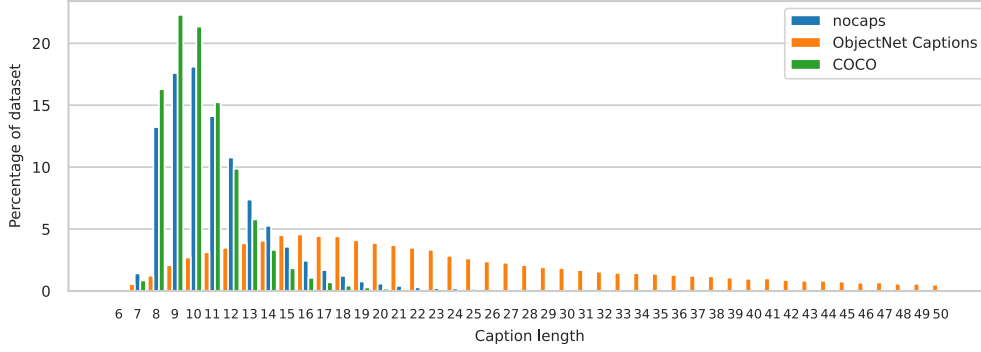


Figure 6: The distribution of caption lengths for ObjectNet Captions compared to those of nocaps and COCO. Note that the graph is cut off at 50 for visibility, although 5108 (5.8%) captions in ObjectNet Captions have more than 50 words. ObjectNet has considerably longer and richer descriptions due to its captions being derived from corrected audio transcriptions.

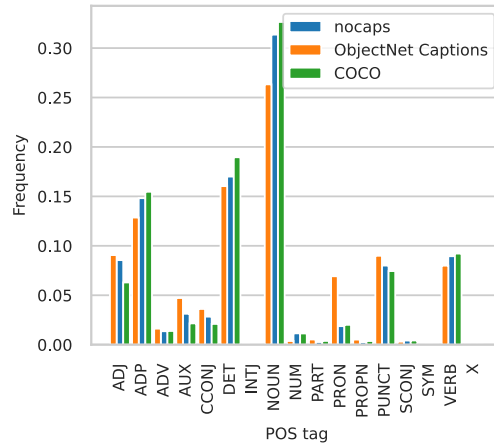


Figure 7: The distribution of part of speech tags for ObjectNet Captions compared to that of nocaps and COCO. Despite the fact that ObjectNet Captions images are of static scenes, the proportion of verbs is similar to that seen in other datasets. ObjectNet Captions has a much higher frequency of pronouns and a somewhat smaller frequency of nouns, among other smaller differences.

| <i>Dependency<br/>relation</i> | <i>ObjectNet<br/>Captions</i> | <i>COCO</i> | <i>nocaps</i> |
|--------------------------------|-------------------------------|-------------|---------------|
| det                            | 16.0                          | 18.9        | 17.0          |
| case                           | 12.2                          | 15.0        | 14.3          |
| punct                          | 9.0                           | 7.4         | 8.0           |
| amod                           | 7.9                           | 6.3         | 8.4           |
| nsubj                          | 6.9                           | 4.2         | 4.6           |
| root                           | 6.3                           | 8.9         | 8.2           |
| nmod                           | 6.0                           | 7.0         | 6.9           |
| compound                       | 5.4                           | 5.0         | 4.6           |
| obl                            | 5.2                           | 7.5         | 6.6           |
| conj                           | 4.1                           | 2.5         | 3.2           |
| cc                             | 3.6                           | 2.1         | 2.8           |
| cop                            | 2.5                           | 0.5         | 0.8           |
| obj                            | 2.4                           | 2.9         | 3.1           |
| acl                            | 1.7                           | 3.5         | 2.9           |
| expl                           | 1.5                           | 0.2         | 0.0           |
| aux                            | 1.5                           | 1.2         | 1.6           |
| advmod                         | 1.5                           | 1.4         | 1.3           |
| acl:relcl                      | 1.4                           | 0.5         | 0.4           |
| compound:prt                   | 0.6                           | 0.5         | 0.4           |
| nmod:poss                      | 0.6                           | 0.7         | 0.7           |

Table 3: Distribution of dependency relations across ObjectNet Captions and two other datasets; shown are percentages. Only the most frequent 20 relations in ObjectNet Captions are shown. ObjectNet Captions stands out in several ways which indicate that the captions are considerably richer by combining together multiple concepts, including: it has twice as many conjunctions, four to five times as many copulas (linking word between a subject and a predicate), three times as many relative clauses.

| Dataset               | Model            | B-1               | B-2               | B-3               | B-4               | R                 | METEOR            | CIDEr              | SPICE             | BERT<br>Score     | CLIP-S            | RefCLIP-S         | HUMANr             |
|-----------------------|------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|-------------------|-------------------|-------------------|-------------------|--------------------|
| COCO                  | GIT <sub>L</sub> | 80.8 ± 0.4        | 66.2 ± 0.5        | <b>52.9</b> ± 0.6 | <b>41.8</b> ± 0.6 | <b>60.3</b> ± 0.4 | <b>30.4</b> ± 0.3 | <b>136.4</b> ± 2.0 | 23.5 ± 0.3        | 71.9 ± 0.1        | 77.3 ± 0.2        | <b>82.9</b> ± 0.1 | -0.05 ± 0.02       |
|                       | ClipCap          | 74.2 ± 0.4        | 57.4 ± 0.5        | 43.2 ± 0.6        | 32.2 ± 0.6        | 55.0 ± 0.4        | 27.1 ± 0.3        | 108.5 ± 1.8        | 20.1 ± 0.2        | 68.9 ± 0.1        | <b>78.3</b> ± 0.2 | 82.6 ± 0.1        | -0.18 ± 0.02       |
|                       | ExpNet           | <b>82.7</b> ± 0.4 | <b>67.7</b> ± 0.4 | <b>53.3</b> ± 0.6 | <b>41.0</b> ± 0.6 | <b>60.3</b> ± 0.4 | <b>30.2</b> ± 0.3 | <b>139.6</b> ± 1.9 | <b>24.4</b> ± 0.2 | <b>73.7</b> ± 0.1 | 76.9 ± 0.2        | 82.7 ± 0.1        | -0.07 ± 0.02       |
|                       | Human            | 63.1 ± 0.4        | 43.5 ± 0.5        | 29.3 ± 0.5        | 19.4 ± 0.5        | 46.5 ± 0.4        | 24.1 ± 0.2        | 87.8 ± 1.5         | 20.8 ± 0.3        | 58.0 ± 0.1        | <b>78.2</b> ± 0.2 | 82.2 ± 0.1        | <b>0.03</b> ± 0.02 |
| nocaps                | GIT <sub>L</sub> | 74.8 ± 0.6        | 61.7 ± 0.7        | <b>48.8</b> ± 0.7 | <b>37.5</b> ± 0.7 | 54.2 ± 0.5        | 25.5 ± 0.3        | <b>94.7</b> ± 1.7  | 12.3 ± 0.2        | 60.6 ± 0.5        | 77.1 ± 0.2        | 82.0 ± 0.2        | -0.25 ± 0.02       |
|                       | ClipCap          | 75.1 ± 0.4        | 57.8 ± 0.5        | 42.1 ± 0.6        | 29.9 ± 0.6        | 52.0 ± 0.3        | 23.8 ± 0.2        | 69.0 ± 1.5         | 10.7 ± 0.2        | 60.1 ± 0.3        | 73.1 ± 0.2        | 77.8 ± 0.2        | -0.37 ± 0.02       |
|                       | ExpNet           | <b>80.3</b> ± 0.4 | <b>64.9</b> ± 0.5 | <b>49.6</b> ± 0.6 | <b>36.6</b> ± 0.6 | <b>55.8</b> ± 0.4 | 25.6 ± 0.2        | 82.2 ± 1.5         | 12.1 ± 0.2        | <b>62.6</b> ± 0.3 | 70.0 ± 0.2        | 76.5 ± 0.2        | -0.35 ± 0.02       |
|                       | Human            | 74.8 ± 0.4        | 56.0 ± 0.5        | 40.3 ± 0.5        | 28.3 ± 0.5        | 52.1 ± 0.4        | <b>27.6</b> ± 0.2 | 86.4 ± 1.5         | <b>15.2</b> ± 0.2 | 58.6 ± 0.3        | <b>78.0</b> ± 0.2 | <b>82.6</b> ± 0.1 | <b>0.01</b> ± 0.02 |
| ObjectNet<br>Captions | GIT <sub>L</sub> | 43.8 ± 0.3        | 33.0 ± 0.3        | 23.4 ± 0.2        | 16.2 ± 0.2        | 36.4 ± 0.2        | 13.3 ± 0.1        | 20.9 ± 0.4         | 8.4 ± 0.1         | 42.1 ± 0.2        | 75.8 ± 0.1        | 77.2 ± 0.1        | -0.46 ± 0.02       |
|                       | ClipCap          | 50.0 ± 0.3        | 35.1 ± 0.2        | 23.2 ± 0.2        | 15.4 ± 0.2        | 35.3 ± 0.1        | 12.4 ± 0.1        | 10.2 ± 0.3         | 6.2 ± 0.1         | 39.7 ± 0.1        | 74.2 ± 0.1        | 73.7 ± 0.1        | -0.69 ± 0.01       |
|                       | ExpNet           | 51.3 ± 0.3        | 38.2 ± 0.2        | <b>26.2</b> ± 0.2 | <b>17.6</b> ± 0.2 | <b>38.5</b> ± 0.1 | 13.9 ± 0.1        | 14.9 ± 0.3         | 8.1 ± 0.1         | <b>43.5</b> ± 0.1 | 72.0 ± 0.1        | 74.4 ± 0.1        | -0.56 ± 0.02       |
|                       | Human            | <b>60.5</b> ± 0.2 | <b>39.9</b> ± 0.2 | 25.4 ± 0.2        | 16.1 ± 0.2        | <b>38.7</b> ± 0.2 | <b>20.4</b> ± 0.1 | <b>31.3</b> ± 0.5  | <b>16.3</b> ± 0.1 | 37.6 ± 0.2        | <b>77.0</b> ± 0.1 | <b>77.9</b> ± 0.1 | <b>0.0</b> ± 0.02  |

Table 4: Full capitulation of evaluation results presented in table 2

**Disclaimer:** This HIT is part of a ██████████ research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. The results of the research may be presented at scientific meetings, published in scientific journals, or made publicly available to other researchers. Clicking on the 'Submit' button indicates that you are at least 18 years of age, you are a native English speaker, and you agree to complete this HIT voluntarily.

**Notice:** If you encounter any issues or find any bugs, please email us **with the copy-pasted text of any error messages you receive**, and we will do our best to fix them. If you consistently encounter this error, please **do not continue to attempt to complete more HITs**, and instead email us with some information about your system configuration including your operating system and web browser version.

**Requirements:** To complete this task, you must be in a relatively quiet environment on a computer equipped with a microphone, using one of the following web browsers: Edge, Chrome, Firefox, Safari, or Opera. **You must have cookies enabled** or you will be unable to submit the HIT.

**Instructions:** You will be submitting audio recordings using the interface below.

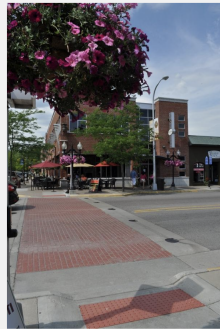
1. When prompted, grant permission to the site to use your microphone for the duration of the HIT.
2. Use the volume meter in the bottom-right of the window to help ensure that your microphone is working properly, and that you are a proper distance away from it. The meter should move as you speak. **If the volume meter does not move, or if the recording button is disabled, please check to make sure that you have given permission to your web browser to access your microphone.**
3. Press the green "Record" button to start recording. After you press it, the button will turn into a red "Stop" button.
4. Complete the task as described below.
5. Press the red "Stop" button to stop recording. After you press it, your audio recording will be processed automatically.
6. If your recording is acceptable, you will be prompted with the next photo. Otherwise, you will be asked to try recording again.
7. Once you have submitted all the necessary recordings, press the green "Submit" button to submit the HIT.

**Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.**

#### (a) Caption recording task consent page

**Task:** Throughout the task, on the left of the screen you will be presented with 4 different images, one at a time. **Please record yourself describing each image as if you were explaining it to someone who could not see it.** We're looking for a couple of sentences per image. You can talk about specific objects, locations, shapes, colors, etc. in the image. For help, refer to the example below.

Here's an example of the level of detail we're looking for:



*"A red brick sidewalk, with a brick building on the far side of the street. There's a patio with umbrellas at the side of the building. In the foreground, there's a hanging plant with pink flowers."*

#### (b) Caption recording example instruction page

**Disclaimer:** This HIT is part of a ████████ research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. The results of the research may be presented at scientific meetings, published in scientific journals, or made publicly available to other researchers. Clicking on the 'Submit' button indicates that you are at least 18 years of age, you are a native English speaker, and you agree to complete this HIT voluntarily.

**Notice:** If you encounter any issues or find any bugs, please email us **with the copy-pasted text of any error messages you receive**, and we will do our best to fix them. If you consistently encounter this error, please **do not continue to attempt to complete more HITs**, and instead email us with some information about your system configuration including your operating system and web browser version.

**Requirements:** To complete this task, you must be able to listen to short audio clips and use a keyboard. You must use one of the following web browsers: Edge, Chrome, Firefox, Safari, or Opera. **You must have cookies enabled** or you will be unable to submit the HIT.

**Instructions:** You will be listening to audio clips and correcting their automatically generated transcripts using the interface below.

1. You may use the "Play example" button to ensure that your audio playback system is working properly. If you click the button and do not hear a sound, please double check your system settings to ensure you can hear audio playback.
2. Press the green "Play" button to begin playback of the sound clip.
3. Use the text box on the left to transcribe the contents of the audio clip. To help you do so, the box is pre-filled with automatically generated text that may contain errors.
4. Press the "Submit" button when you have finished correcting the transcript.
5. Once you have submitted all the necessary transcripts, press the green "Submit" button to submit the HIT.

**Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.**

### (a) Transcript task consent page

**Task:** Use the green "Play" button to listen to an audio clip. Using the text box on the left, correct the audio transcription. You should fix incorrectly transcribed words and add punctuation, but don't add words that are not in the recording. When you are done, click the green "Submit" button. For help, refer to the example on the right.

Transcript 1 of 4

Here's an example of how to correct a transcript:

**Play example**

**Before:** "a bluish cylinder wedged between a toaster and a microwave the cylinder is resting on a wooden surface and is possibly an exercise device"

**After:** "A bluish cylinder wedged between a toaster and a microwave. The cylinder is resting on a wooden surface and is possibly an exercise device."

Play

A Febreze bottle on a brown wooden dresser next to some Benzomatic.

Submit

### (b) Transcript task instruction page

### Informed consent to participate in this study

This HIT is part of a ██████████ research project. Your decision to complete this HIT is voluntary. There is no way for us to identify you.

The only information we will have, in addition to your responses, is the time at which you completed the survey and generic non-identifiable about your computer such as its resolution and browser version number.

The results of the research may be presented at scientific meetings or published in scientific journals.

The responses collected in this experiment will be released to the scientific community and the public.

Clicking on the 'SUBMIT' button on the bottom of this page indicates that you are at least 18 years of age and agree to complete this HIT voluntarily.

☐ I agree

(a) HUMANr task instruction page

### Instructions

PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY BEFORE BEGINNING:

You are about to be shown a series of images. Please look closely at all parts of the image.

For each image there will also be a pair of description written in text boxes below the image. Your task is to pick which description best matches the image on a scale from 1 to 9. 1 means only the description on the left could describe the image. 9 means only the description on the right could describe the image. 5 means that both descriptions match the image equally well.

Please read the captions carefully and compare them with your image. Your performance will be recorded and analyzed. If you do not take the task seriously we will know and will have to reject your work.

Thank you!

(b) HUMANr task instruction page