

# VICR Datasheet

<b>Motivation.....</b>	<b>1</b>
<b>Composition.....</b>	<b>2</b>
<b>Collection Process.....</b>	<b>4</b>
<b>Preprocessing/cleaning/labeling.....</b>	<b>5</b>
<b>Uses.....</b>	<b>5</b>
<b>Distribution.....</b>	<b>6</b>
<b>Maintenance.....</b>	<b>7</b>

## Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on assessing the quality of a caption for an image.

**Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset curation was the work of several scholars from San Francisco State University involving graduate and undergraduate students, coordinated by Prof. Ilmi Yoon.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided by the Smith-Kettlewell Eye Research Institute and also by Ability Central and a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90 REGE0018).

**Any other comments?**

None.

## Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is an image-caption pair. It includes the URL of an image, a corresponding caption that may or may not accurately describe the image, and a list of between 3 and 7 whole number ratings on a scale from 1 to 5 indicating the quality of the caption for the image.

**How many instances are there in total (of each type, if appropriate)?**

There are 15,646 image-caption pairs in total. There are 9,990 unique images, most of which are associated with a single caption, but some are associated with as many as 10 captions.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The images in the dataset came from the MS-COCO 2014 validation set and the Flickr8k-Expert dataset. The captions were selected from five sources: (1) the original MS-COCO captions, (2) captions generated using the Pythia framework, (3) captions generated using the GLACNet model, (4) mismatched MS-COCO captions from other images, and (5) the original Flickr8k-Expert captions. The captions have an average length of 10.9 words, where the shortest is 2 words, and the longest is 30 words.

**What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance has the URL of an image, the caption for the image as plain text, the subset (train, test, or val) of the pair, the number of ratings associated with the pair, and the list of individual ratings from 1 to 5 of the quality.

**Is there a label or target associated with each instance?** If so, please provide a description.

The list of ratings for the caption quality is the label for the image-caption pair. In practice, each individual rating can be used as a label for the image-caption pair, or alternatively they could be aggregated as a single target average for the pair.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Not all image-caption pairs have the same number of ratings, but all have at least 3 ratings, and no other data is missing.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

Certain images appear in multiple instances, and some captions are associated with multiple images.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

We have provided the recommended data split as a column in the dataset. We ensured no image appears in more than one split, but captions were not checked for overlap.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

Due to the subjective nature of the task, there is no absolute ground truth, so the ratings are inherently noisy.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The images are provided as URLs but the rest is self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

No.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.**

No.

## Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The ratings were collected through the use of our game. Participants clicked buttons on a webpage and the ratings were saved to a database.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**  
How were these mechanisms or procedures validated?

A web-based game was used to collect user ratings. It was designed to only allow users to submit valid responses.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

Some of the images were randomly sampled from MS-COCO 2014, but the image-caption pairs and their associated ratings are not themselves sampled from a larger dataset.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Students were recruited via departmental mailing list to participate in the data collection game. Participants were rewarded monetarily based on their score, earning \$24 per hour on average.

**Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected from two data collection runs. The first was conducted over roughly 5 weeks with a break, during the summer and fall of 2021 (July 31st, 2021 through August 20th, 2021, and August 31st through September 16th, 2021). The second was conducted over 10 days, during the fall of 2022, (November 3rd, 2022 through November 14th, 2022).

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, an IRB was consulted, and the protocol was approved. Information about San Francisco State University's Institutional Review Board may be found on their homepage at [https://research.sfsu.edu/hap\\_IRB](https://research.sfsu.edu/hap_IRB).

## Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

No.

## Uses

**Has the dataset been used for any tasks already?** If so, please provide a description.

The dataset has been used to train a machine learning regression model to automatically rate the quality of unseen captions for images.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

<https://ai.youdescribe.org>

**What (other) tasks could the dataset be used for?**

It could be used for rating captions given an image, or also rating an image given a caption.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

No.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

No.

**Any other comments?**

None.

## Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

Yes, we intend to make the dataset publicly available.

**How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

It will be available through GitHub and our website <https://ai.youdescribe.org> as a csv, tarball, and zip.

**When will the dataset be distributed?**

Fall 2023.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be licensed under a Creative Commons CC-BY-NC-SA 4.0 license.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

**Any other comments?**

None.

## Maintenance

**Who will be supporting/hosting/maintaining the dataset?**

The authors will be supporting/hosting/maintaining the dataset.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

Through our website <https://ai.youdescribe.org>.

**Is there an erratum?** If so, please provide a link or other access point.

Not yet.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

We do not plan to update this dataset.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

Not applicable.

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

No.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

There is no such mechanism at the moment.

**Any other comments?**

None.