Validated Image Caption Rating (VICR) Scale, Dataset, and Model

Anonymous ACL submission

Abstract

Assessing the quality of an image caption is a complex task. We propose a new image caption rating system that consists of (1) a robust rating scale that is consistent, teachable, and externally validated, (2) an engaging and scalable data generation approach for the task, (3) a high-quality dataset, and (4) an effective image caption rating predictor. Using contemporary approaches from psychometrics we demonstrate that the proposed scale and rater training routine can support high quality annotation efforts for the task. We introduce two new datasets (one original and another derived) for the task. Our reference-free and multi-level rating predictor performance is on par with state-of-the-art approaches.

1 Introduction

We present a novel image caption rating (ICR) framework that consists of (1) externally validated rating scale, (2) a scalable data generation tool, and (3) high-quality dataset, and (4) an effective ICR prediction model. The problem of image caption quality estimation has received substantial attention in recent years, underscoring the increasing need for reliable solutions (Jiang et al., 2019; Lee et al., 2021; Hessel et al., 2021; Lee et al., 2020; Wang et al., 2018). Existing datasets for the task of image caption rating are generated using the traditional approach of human-driven data annotation efforts, and typically use ad hoc rating scales (Levinboim et al., 2019; Hodosh et al., 2013; Vedantam et al., 2015). All these datasets have been tremendously valuable in advancing the field and have been used extensively (Hessel et al., 2021; Lee et al., 2021). However, several of the datasets suffer from high skew in ratings and mixed quality annotations. Our work seeks to improve the rigor, quality, and scalability of ICR datasets and data generation process, and provides a robust scoring instrument that is informed by contemporary approaches to measurement – specifically, Item-Response Modeling.

For the problem of image caption rating estimation, the main difference in existing approaches stems from their ability to estimate the rating in the presence or absence of reference caption(s). BLEU (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2014), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016), BERTscore (Zhang et al., 2019) and ViLBERTscore (Lee et al., 2020) belong to the former category where reference captions are essential, while Visual Semantic Embedding Plus Plus (VSEPP) (Faghri et al., 2017), CLIPScore (Hessel et al., 2021) and approaches proposed by Cui and colleagues (2018) and Levinboim and colleagues (2019) can operate without reference captions. The ability of these approaches to assess caption quality without requiring reference captions has led to rapid progress on this problem. However, the rating granularity employed by these approaches has been restricted to simple binary scale (*good* or *bad* caption). In our work we seek to lift this restriction by employing a 5-level rating scale that can model different aspects of quality in the context of image captions (e.g., correctness, completeness, and inclusion of local and global context), while also retaining the benefits of reference-free rating approach. Although a more detailed scale can offer higher rating capacity, it can also increase the complexity of the rating task; potentially making the task more subjective and tedious. To tackle this downside, we propose a two-pronged solution during data generation: (1) rigorous training procedure with in-built quality control, and (2) gamification.

Altogether we refer to our work as a Validated Image Caption Rating (VICR) framework; and our specific contributions are introduction of the VICR Scale, VICR Game, VICR Dataset, and VICR Model. The rest of the paper is organized as follows. The next section provides the context and the rich prior work that we build our work on. Section 3 describes the VICR system in detail, followed by Results and Analysis in Section 4, and the conclusions we draw from this work in Section 5.

2 Related Work

2.1 ICR Scale and Datasets

Google Image Caption (GIC) Dataset (Levinboim et al., 2019) and Flickr8k-Expert (Hodosh et al., 2013) are the two widely used large image caption datasets that also include ratings. GIC dataset has 600K image-caption ratings. For each image/caption pair, 8-10 binary ratings were collected. The ratio of good ratings to total ratings is used as image caption quality score (range: [0, 1]). As is common with binary scales, it does not have the capacity to handle incomplete or partially correct captions. Figure 1 includes two illustrative examples.

Similarly, Conceptual Caption Challenge human evaluation studies on the T2 test dataset[1] contains 5000 image and caption pairs and the human ratings are collected in the same manner of GIC; each pair has the total rating counts and the good ratings counts.

Caption: “a slab of beef in a baking dish”	Caption: “the boat in the water”

For both images all 10 raters chose “good” rating even though salient aspects of the image are not captured by the caption (shoe shape in the beef slab over the cutting board; the docked nature of the ship).

Figure 1: Two examples from Google Image Caption Dataset illustrating the limitation of binary scale.

Flickr8k-Expert, a subset of Flickr8k dataset (8,000 images and 5 captions per image), has 5,822 captions across 1,000 images where each caption has received 3+ ratings from human annotators (21 college students). The rating scale used for Flickr8k dataset consisted of 4 levels (Table 1). The complexity of the ICR task combined with the underspecified rating scale and human error lead to fairly low inter-rater agreement. The rating distribution is also heavily skewed toward levels 1 & 2, indicating overall lower caption quality.

r	**Meaning**
4	Describes the image without any errors.
3	Describes the image with minor errors.
2	Is somewhat related to the image.
1	Is unrelated to the image.

Table 1: Flickr8k-Expert ratings and meanings.

The CapEval1K dataset (Lee et al., 2021) is rich for containing fluency, relevance, and descriptiveness rates per caption, but has a rather small size (1,000 captions for 250 images). The PASCAL50s dataset (Vedantam et al., 2015) has 50 reference captions per image for 1000 images, but the ratings are not in numeric scale.

2.2 Reference-free ICR Estimators

VSEPP (Faghri et al., 2017) and CLIPScore (Hessel et al., 2021) are multimodal, visual-linguistic models that use cosine similarity to measure the distance between an image embedding and text embedding in a shared visual-semantic embedding space. Unfortunately, while the cosine similarity does a good job on approximation of the similarity of the vectors in the shared visual-linguistic semantic space, fine tuning or manipulation of the similarity of the image and language features remains difficult.

Cui and colleagues (2018) created a deep learning method for determining if a caption for an image was human-written or machine generated. However, this is a binary classifier and is not sufficient for diverse use cases.

Levinboim and colleagues (2019) developed an image-caption Quality Estimation (QE) model by training a deep learning model on the GIC dataset. The model inherits the limitations from the dataset discussed in Figure 1.

[1] https://www.conceptualcaptions.com/winners-and-data

Lee and colleagues (Lee et al., 2021) developed Unreferenced Metric for Image Captioning (UMIC) using UNITER (Chen et al., 2020) via contrastive learning, a process where the model is trained to compare and discriminate the ground-truth captions and diverse synthetic negative samples. Jiang and colleagues (Jiang et al., 2019) developed TIGEr (Text-to-Image Grounding for Image Caption Evaluation by improving the mapping of the image and the caption pair into carefully grounded vector spaces. These approaches improved consistency with human judgements over prior metrics, but still did not exceed .5 Kendall τ scores on the Flickr8K expert data set.

2.3 Integrative Inferential Reasoning (IIR)

The importance of a robust rating scale for the ICR task cannot be overstated. Having a theoretical foundation can ensure that a rating scale yields explicit and trainable scoring guides that lead to reliable ratings. Based on industry-accepted image description guidelines, the context in the image must be included in the caption, and thus is an important aspect of caption quality (Rai et al. 2010)[2]. Contextual integration is the backbone of Integrative-Inferential Reasoning (IIR) (Blum et al., 2020). IIR is a cognitive framework that structures context integration in text- and image-based narratives. IIR's scaled definitions of context and inference offers a roadmap for training humans (and by extension, machines) on how to rate image caption quality based on these characteristics. In its modern form, IIR is a novel approach to capturing combined notions of context and inference; however, the theory stems from older notions of local (e.g., propositional or literal) and global (e.g. schematically or culturally relevant) coherence, which has been investigated in literacy (Graesser et al. 1994; Language and Reading Research Consort...), cognition (Frith and Happé 1994; Van der Hallen et al. 2015), neurodiverse populations such as autism (Happé & Frith, 2006; Nuske & Bavin, 2011); and the schema of Question-Answer Relations (Pearson and Johnson 1978; Raphael and Au 2005). With its historical theoretical grounding, IIR offers an exciting foundation for developing a new kind of image rating scale.

3 Methods

The old adage "A picture is worth a thousand words." perfectly captures the challenge faced by image caption raters (humans and machines). An image can convey layers of nuanced information, while a short textual caption has a very limited information bandwidth. Naturally, assessing the quality of image captions is an inherently tricky task. To tackle this complex problem, we start by unpacking the ICR pipeline. The first source of error is often the rating scale itself. The errors caused by an ill-defined scale propagate downstream and compound. The second source of error is typically humans who are doing the tedious and complicated task of rating the captions. We coalesce these observations to define two key objectives for our work:

Objective #1: Design and develop a reliable and scalable data generation approach for the task of image-caption rating. To achieve this objective, we innovate along three areas: (1) Develop a rating scale that accurately captures the nuances and aspects of image caption quality (VICR Scale); (2) Develop an engaging tool (VICR Game) to facilitate high-quality data generation from human raters (VICR Dataset); and (3) Assess the ability of human raters to effectively use this data-generation approach.

Objective #2: Develop a novel image-caption rating model (VICR Model) that employs the outcomes from objective #1.

Together, these objectives provide a robust, high-quality, and scalable image-caption rating system which is described next.

3.1 VICR Scale: Relating IIR to Image Captions

Integrative Inferential Reasoning (IIR) is a theoretical construct, developed using the BEAR assessment system (Wilson, 2005). We applied IIR as a theoretical foundation (IC-IIR) to inform the development of VICR Scale. This 5-level scale captures nuances in caption accuracy, completeness, inferential, and contextual information as listed in Table 2.

To evaluate the efficacy of the VICR Scale at training raters and at producing consistent ratings across raters, we employed measures of rater competency using the following approach. Rater competency was represented by "items" (the image-caption pairs) and "responses"

[2] https://dcmp.org/learn/descriptionkey

(participants' ratings). We used a 5x5 factorial items design: five images were used, and each image was paired five times, with captions representing each of the five levels of the VICR Scale. Each rater was assigned a *competency score* based on the degree of agreement between their ratings and expert ratings as follows: *Exact Agreement* (participant and expert ratings are equal) received a score of 2; *Adjacent Agreement*, (participant and expert ratings differ by 1) received a score of 1, and *Lack of Agreement* (participant and expert ratings differ by more than 1) received a score of 0. The cumulative score over all 25 image-caption pairs was computed for each rater and analyzed using the Partial Credit Model (PCM) (Masters, 1988; Masters, 2016). The PCM is a Rasch-family measurement model that is used to place items and participants on the same scale and evaluate the quality of an obtained measurement. We also used a Latent Regression (Wilson & De Boeck, 2004) to regress rater competency on their tutorial score obtained during training. Results of these analyses are shared in Section 4.1

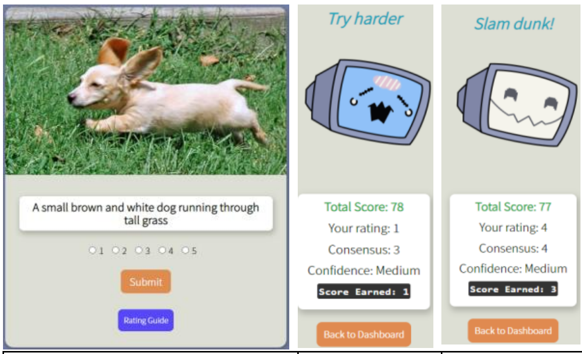
r	**Meaning**
5	Objects, a general scene, and actions are correctly identified if present in the image. The caption describes what is seen and where things are in space.
4	Objects and/or a general scene and/or an action are correctly identified but not every element is completely identified. The caption describes what is seen and where things are in space. There is no interpretation of an event.
3	Relevant objects are correctly identified. The caption describes what is seen but not where objects are in space. There is no description of the overall setting and no interpretation of an event.
2	Objects are partially correctly identified with some errors, but the caption is accurate enough to give an idea of what is happening in the image. The caption identifies most of the objects but might not identify everything. There is no interpretation of what anything means.
1	Objects are incorrectly identified. The caption gives the wrong idea about what is happening in the image.

Table 2: VICR Scale: Ratings and Meanings.

3.2 VICR Dataset Generation: Image Caption Rating Game

To facilitate generation of high-quality and substantially sized data we focus on human rater training and engagement in this phase of the VICR system. Rater training is essential for any data annotation effort, but it is especially important in our project due to the detailed nature of the VICR Scale. A 5-level scale with each level capturing multiple aspects of caption quality is non-trivial to apply for most humans.

Rater Training: The training is conducted online through a web application that starts by showing the VICR Scale to the human rater. When ready the rater proceeds to a test round where an image and caption pair is displayed, and the rater has to choose the most appropriate rating level from the VICR Scale for the pair. This is repeated for 20 image-caption pairs. The accuracy of the rater's selections is computed using the ground-truth ratings. Raters with accuracy of 0.5 or higher are cleared for data generation, and others are required to redo the training until minimum accuracy is met. The reasoning behind the chosen accuracy threshold is explained in Section 4.1.



(a) Image, caption and 5-level scale. VICR Scale can be consulted anytime through the button.	(b) Feedback when score earned is 1	(c) Feedback when score earned is 3.

Figure 2: VICR Image Caption Rating Game.

VICR Game: To promote rater engagement we frame the annotation task as a single-player asynchronous competitive game; following on the path of image labeling ESP game (Von Ahn & Dabbish, 2004). The web-based VICR Game is designed to provide a similar user experience as the training phase – an image-caption pair is displayed, and the player selects the appropriate rating from the 5-level VICR Scale (Fig. 2a). After rating submission, the player receives feedback that compares their selection with those of the other players so far (Fig. 2b and 2c). Specifically, a *consensus score* (*con*), which is the rounded average of all the previous ratings for that image-caption pair so far, is displayed. Player earns points, p, for the rating submission using the following formula:

$$p = \max(4 - \lceil 2\,|x - con| \,/\, c \rceil, -1)$$
$$x = \text{player's rating selection}$$
$$con = \text{consensus score}$$
$$c = 1 + (1 + (n - 1)\, \sigma^2 \,/\, V_{\max}) \,/\, n$$

where σ^2 is the variance of the previous ratings and $V_{\max} = 4$, the largest possible variance of the previous ratings, therefore $p \in [-1, 3]$. This formulation models two intuitions: 1. the assigned points should be inversely proportional to the difference between the player's rating and the consensus score, and 2. the assigned points should be proportional to the degree of agreement among the ratings so far. Together, these intuitions ensure low points for scenarios where agreement among prior ratings is high and the current rating exhibits a large difference from the average. In contrast, if the level of agreement is low, the points decrease only gradually as the difference from the average increases. This is supported by the coefficient c in the formulas above. This coefficient, called confidence, will be between 1 and 2, where 1 represents perfect confidence, and 2 represents the least possible confidence. It is used to modify the distance from the consensus at which various points are awarded.

This formulation provides the ability to penalize ratings that deviate substantially, $p \in [-1, 3]$. We seed the target ratings initially with ratings from VSEPP. For the purposes of calculating mean, variance, and the level of consensus multiplier, we include this initial rating twice, i.e., as two agreeing data points. Once a participant gives their rating for the image-caption pair, this rating replaces one of the two initial ratings, and once a second player has rated the pair, the second initial rating is replaced as well, so that the average and level of consensus are now purely based on the two human ratings, and from then on, the human ratings accumulate as normal.

3.3 VICR Model: Image Caption Rater

We propose a multi-level reference-free image-caption rating predictor, VICR Model (Fig. 3). The rating predictor starts by converting the image and the caption into image and language embeddings, respectively. Preliminary experiments with various image and language embeddings, demonstrated ViLBERT co-fusion embeddings as being the most effective for our model. We use the pooled text and image embeddings of the final hidden layer in ViLBERT and concatenate these into a 2048-dimensional vector as input to our network. For the regressor model, a two-hidden-layer fully-connected neural network with 512 neurons on the first layer, followed by ReLU activation, and 256 neurons on the second layer, followed by another ReLU activation, with a single neuron with linear activation as the output layer. We used 80% dropout on both hidden layers.

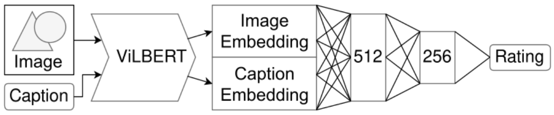


Figure 3: Schematic diagram of the VICR model architecture.

4 Experiments and Results

Our evaluation methodology for the VICR system consists of two user studies for VICR Scale and VICR Game, respectively, a comparative analysis of VICR Dataset, and an empirical evaluation for the VICR Model.

4.1 VICR Scale: Initial Validation of Image-Caption Quality and Rater Consistency

The goal of User Study 1 was to evaluate the efficacy of a VICR Scale and the corresponding training module at generating high-quality ICR data. For this study, 132 fully anonymized participants (college students at a 4-year public university) were recruited. The participants started by undergoing the Rater Training routine (Section 3.2), and the ones who cleared the quality threshold were then prompted to rate 25 items (5 images and 5 captions per image in random order) using the 5-level VICR Scale.

The collected data was then analyzed as per the methodology described in Section 3.1. Specifically, we employed Wright Map – an analytical tool that allows us to place human raters and image caption pairs (i.e., items) visually on the same scale (Embretson 1996; Stachl and Baranger 2020; Blum et al. 2020; Wilson 2005), (Brondfield et al. 2021; Blum 2019) to analyze rater competency when using VICR Scale, and a latent regression analysis understand the strength of an explanatory relationship between our training module and rater competency.

4.1.1 Wright Map

The PCM (section 3.1) uses human rater proficiency and image-caption pair difficulty estimates, and the error associated with them to generate Wright Maps (also known as item-person maps, Fig. 4). The first column displays results from a latent regression (section 4.1.3); the second column shows a histogram of rater competency scores (Section 3.1) in logits (column three).

The key observation from this analysis comes from the right side of the Wright Map which reports participants' levels of VICR Competency. Two cumulative thresholds per item (image-caption pair) are represented on the right side (25 columns, one for each item along the horizontal axis). The first threshold, marked in yellow, represents where a respondent would be equally likely to score 0 (*LA: Lack of Agreement*) vs. 1 or 2 (*AA: Adjacent* or EA: *Exact Agreement*); the second threshold, marked in red, represents where a respondent would be equally likely to score 0 or 1 (*LA* or *AA*) vs. 2 (*EA*). This Wright Map shows that in our data most of the items' second thresholds are above most of the items' first thresholds, which represents internal validity of raters' competency in using the VICR Scale (no/minimal confusion).

4.1.2 Latent Regression

Latent regression was used to explain the relationship between VICR training routine and rater competency. The regression coefficient for *training score* was 0.17 (stderr 0.03) which is significant at the .05 level. This means that each additional tutorial item that the respondent rated correctly is associated with a mean increase in rater competency of 0.17 logits. The significance of this output is seen in the leftmost column of the Wright Map, which shows the predicted mean VICR competency score for each possible tutorial score from 5 to 10. At the training score of 5, the predicted mean rater competency (-0.13) is well above all the first thresholds, and above the second threshold for 10 of the 25 items. This suggests that, on average, even a respondent with training score of 5 (weakest rater) is very likely to demonstrate at least *AA* on all items and has more than 50% chance of *EA* on 10 of the items (and less than a 50% chance for *EA* on the other 15 items). The significant finding from this analysis is that it provides external validity of raters' competency in using the VICR Scale. This analysis also informs the choice of minimum threshold (training score of 5 = 0.5 accuracy) used for rater selection during training routine (Section 3.2).

4.1.3 Is the VICR Scale teachable?

The short answer is, yes. The positive and significant regression coefficient of 0.17 indicates that respondents who were more successfully trained in using the VICR Scale were better able to reach *EA* on more image-caption pairs. This is also visible in the Wright Map, where raters with an increasing training score of 5 to 10 are more likely to reach *EA* on up to 80% of the items, and most likely to have *AA* on 100% of the items with a training score of 10.

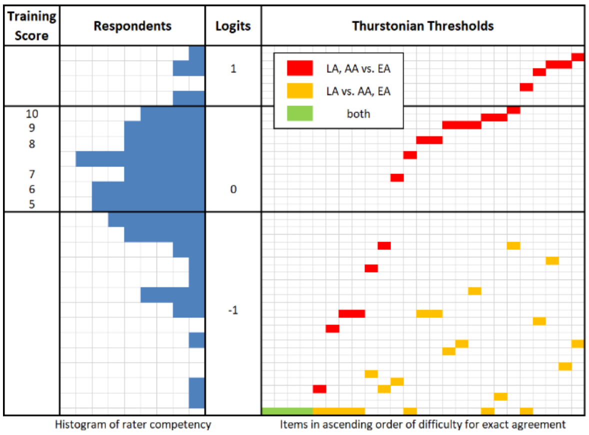


Figure 4: Wright Map augmented with predicted means from latent regression

4.1.4 Can people use the VICR Scale consistently and well?

Based on respondent frequency and locations on the Wright Map, most respondents (103 out of 132, i.e., 78%) are above all the first thresholds; as noted above, respondents at that level can reliably achieve high agreement (*EA* or *AA*) (median was 0.01 logits). Respondents who were more successfully trained (higher rater competency score) tend to achieve *EA* on more of the items; respondents with a very high training score (10) tend to achieve *EA* on 80% of the image-caption pairs.

4.2 VICR Game and Datasets

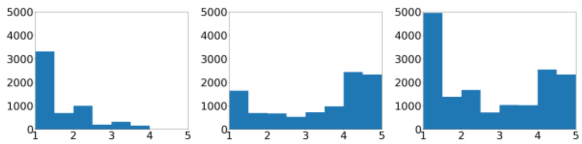
The goal of User Study 2 was to employ the VICR Game to generate a new dataset for the ICR task. We also created the Combined Dataset that consists of the new VICR Dataset and the Flickr8k-Expert dataset. (The new datasets will be freely available to the research community.)

VICR Dataset: A collection of image-caption pairs was assembled for this user study as follows: 8,990 distinct images were chosen at random from MS-COCO 2014 Validation subset (Lin et al., 2014), The captions were selected from 4 sources: (1) the original MS-COCO caption, (2) generated using the Pythia framework (Jiang et al., 2018), (3) generated using the GLACNet model (Kim et al., 2018), and (4) mismatched captions from other images. Leading to 9,982 image-caption pairs.

As part of the user study 72 participants (college students) played the VICR Game to rate image-caption pairs from the above collection. On average, the participants played for 102 minutes, earning about $15 per hour. The participants took about 10 seconds on average to rate an image caption pair. By the end of the user study, a total of 48,174 ratings were collected, so that each of the 9,982 image-caption pairs had at least 4 and at most 7 ratings.

Combined Dataset: We also made a Combined dataset composed of Flickr8k-Expert and VICR to create a bigger data set (15,804 image-caption pairs with ratings). When consolidating the two datasets, we mapped Flickr8k-Expert's 4-level Scale to the first 4 levels of VICR Scale, since the meanings ratings to 1 to 5 but instead kept them as their original scale of 1 to 4 since their 1 to 4 map to our 1 to 4 relatively well with 5 being an extra level in our dataset. The 5th level is essentially not represented in the Flickr8k-Expert rating scale.

Comparative Analysis: The rating distribution of the Flick8k-Expert, VICR, and Combined Datasets are illustrated in Fig. 5. For each image-caption pair, the rounded average of all available ratings for that pair is used as the single value rating for the pair.



(a) Flickr8k -Expert (5,664 datapoints) (b) VICR (9,982 datapoints) (c) Combined (15,804 datapoints)

Figure 5: Datasets: Rating Distributions

Figure 5 demonstrates that the new VICR Dataset is less skewed in its rating distribution than the Flickr8k-Expert dataset. It does however exhibit bimodal distribution indicating a larger proportion of low- and high-quality captions than average quality captions. The Combined Dataset naturally embodies the properties of both the source datasets.

4.3 VICR Model

We evaluate the effectiveness of our multi-level reference-free image-caption rating predictor, VICR Model, with two empirical experiments. **Experiment 1:** Table 3 provides results for the first experiment where $VICR_{VICR}$ (VICR model trained with VICR Dataset) performance is compared to Reference-based and Reference-free approaches. (We used Adam optimization, minimized on MSE, for 4,000 epochs.)

It is not surprising that the Reference-based approaches exhibit higher performance. Within the Reference-free category, CLIPScore provides the highest performance, with $VICR_{VICR}$ being a close second. VICR Model shows good generalizability – despite being trained on VICR Dataset with 5-level rating scale, the predictor provides competitive performance on Flickr8k-Expert dataset with 4-level scale.

Reference-based Approaches	τ_C
BLEU-1	36.3
BLEU-4	33.1
METEOR	43.6
ROUGE	38.1
CIDER	43.7
SPICE	45.9
RefCLIPScore	52.7
ViLBERTScore-F	54.2*
Yi et al.	48.1*
Reference-free Approaches	τ_C
CLIPScore	51.5
UMIC-c	43.1*
TIGEr	49.3*
VSEPP	48.1
VisualEntailment	44.6
$VICR_{VICR}$	50.9

Table 3: Kendall τ correlation with ground truth ratings on Flickr8k-Expert dataset for various metrics and predictors. We recreated all the listed results except for the ones with * which are directly from the respective papers. We used "method A" in aggregation (Hessel et al., 2021) and τ_C to be consistent with prior work.

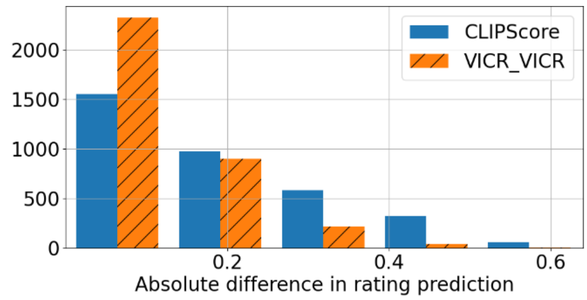


Figure 6: Error analysis: Histogram of absolute error in rating predictions.

Figure 6 provides a deeper analysis of rating predictions by computing the absolute difference between the predicted rating and the ground truth rating for image-caption pairs in the Flickr8k-

Expert dataset. The ratings have been normalized into the range [0, 1]. The x-axis specifies the absolute error in rating prediction. Notice that the 0-error bar for $\text{VICR}_{\text{VICR}}$ is substantially higher than that of CLIPScore. Overall, the histogram distribution for $\text{VICR}_{\text{VICR}}$ is heavily skewed to the left, indicating lower incidence and magnitude of prediction errors.

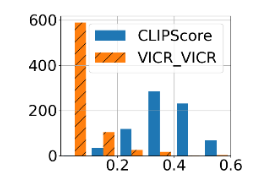
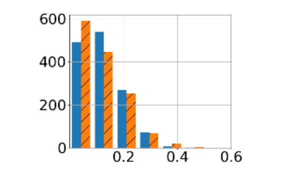
	
(a) Histogram of absolute error in rating predictions of labels in the range [0.0, 0.2]	(b) Histogram of absolute error in rating predictions of labels in the range [0.8, 1.0]

Figure 7: Error analysis: Histogram of absolute error in rating predictions on sub-ranges of labels.

We further analyzed absolute errors in rating predictions on sub-ranges of ground truth ratings (2 shown in Fig. 7), showing higher performance over CLIPScore in almost all ranges.

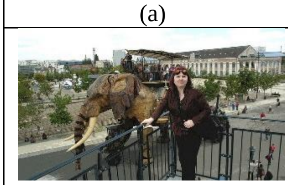
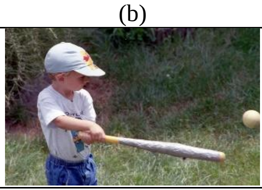
	
Caption: A group of elephants by some buildings on the water.	Caption: a number of baseball players in a field
CLIPScore: 0.43 $\text{VICR}_{\text{VICR}}$: 0.0 human rating avg: 0.0 (a)	CLIPScore: 0.44 $\text{VICR}_{\text{VICR}}$: 0.06 human rating avg: 0.06 (b)
	
Caption: A woman standing on a balcony in front of an elephant float.	Caption: A woman preparing to hit a tennis ball while a man watches.
CLIPScore: 0.69 $\text{VICR}_{\text{VICR}}$: 0.81 human rating avg: 0.81 (c)	CLIPScore: 0.54 $\text{VICR}_{\text{VICR}}$: 0.02 human rating avg: 0.43 (d)

Table 4: Samples of image, caption and metrics.

The examples in Table 4 (from VICR Test set) illuminate this further. For easier comparison, the ratings are all normalized to lie in the range [0, 1]. There are cases where VICR scores align perfectly or very closely with human ratings (e.g., Table 5-a, b, c). There are also cases where $\text{VICR}_{\text{VICR}}$ seems even more accurate than human ratings (e.g., examples d).

Experiment 2: The second experiment studies the ability of the three datasets (Flickr8k-Expert, VICR, and Combined) at training an effective rating predictor with VICR Model. Each dataset was split into 64% training, 16% validation, and 20% test for this experiment. Three models, $\text{VICR}_{\text{Flickr8k}}$, $\text{VICR}_{\text{VICR}}$, $\text{VICR}_{\text{Combined}}$, were trained on the respective Training sets.

Reference-free Approaches	τ_C Flick8K-Expert	τ_C VICR	τ_C Combined
$\text{VICR}_{\text{Flickr8k}}$	52.1*	61.3*	71.2*
$\text{VICR}_{\text{VICR}}$	50.6*	**66.4***	73.4*
$\text{VICR}_{\text{Combined}}$	**53.2***	66.0*	**75.5***
CLIPScore	(51.2[1]) 51.5	66.3	65.7
VSEPP	48.1	62.3	66.5
VisualEntailment	44.6	54.6	65.0

Table 5: Kendall τ correlation with ground truth ratings for reference-free approaches, *Calculated on Test set of each dataset. [1]Reported in (Hessel et al., 2021)

The top half of the Table 5 reports performance of the VICR models with the three Test sets. All three models perform better on VICR and Combined Datasets. This trend is also seen with the other Reference-free approaches (lower half of Table 5). This suggests that the VICR and Combined are more reliable ICR datasets than Flickr8k-Expert.

5 Conclusions

In this work we introduced an image caption rating system that consists of a new rating scale, an engaging data generation approach, a high-quality dataset, and a rating prediction model. A multi-level rating scale that captures various nuances of caption quality can be difficult to apply. Our user studies suggest that a well-defined scale along with methodical training and a game-based data generation setup can provide the right balance of data quality and quantity. The new dataset generated by this approach when employed to train our reference-free rating predictor provides one of the highest performances for the image caption rating task. However, we have not yet explored how to utilize specific context in rating scale or how robustly it performs on objects it is not trained on. We also have not explored potential risks of biases in image caption ratings.

References

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016, October). Spice: Semantic propositional image caption evaluation. In *European conference on computer vision* (pp. 382-398). Springer, Cham.

Blum, A. M. (2019). *Deficit or Difference? Assessing Narrative Comprehension in Autistic and Typically Developing Individuals: Comic vs. Text*. University of California, Berkeley.

Blum, A. M., Mason, J. M., Kim, J., & Pearson, P. D. (2020). Modeling question-answer relations: the development of the integrative inferential reasoning comic assessment. *Reading and Writing*. https://doi.org/10.1007/s11145-020-10026-4.

Brondfield, S., Seol, A., Hyland, K., Teherani, A., & Hsu, G. (2019). Integrating Concept Maps into a Medical Student Oncology Curriculum. *Journal of Cancer Education*, 1-7.

Chen, Yen-Chun et al. "UNITER: UNiversal Image-TExt Representation Learning." Lecture Notes in Computer Science (2020): 104–120. Crossref. Web.

Cui, Y., Yang, G., Veit, A., Huang, X., & Belongie, S. (2018). Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5804-5812).

Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376-380).

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201-212.

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Frith, U., & Happé, F. (1994). Autism: beyond "theory of mind." *Cognition*, *50*(1), 115–132.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36*(1), 5–25.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853-899.

Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., ... & Gao, J. (2019). TIGEr: text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., & Parikh, D. (2018). Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Kim, T., Heo, M. O., Son, S., Park, K. W., & Zhang, B. T. (2018). Glac net: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*.

Language and Reading Research Consortium, & Muijselaar, M. M. L. (2018). The dimensionality of inference making: Are local and global inferences distinguishable? *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, *22*(2), 117–136.

Lee, H., Yoon, S., Dernoncourt, F., Kim, D. S., Bui, T., & Jung, K. (2020, November). ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 34-39).

Lee, H., Yoon, S., Dernoncourt, F., Bui, T., & Jung, K. (2021). UMIC: An unreferenced metric for image captioning via contrastive learning. *arXiv preprint arXiv:2106.14019*.

Levinboim, T., Thapliyal, A., Sharma, P., & Soricut, R. (2019). Quality estimation for image captions based on large-scale human evaluations. *arXiv preprint arXiv:1909.03396*.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.

Masters, G. N. (1988). The Analysis of Partial Credit Scoring. *Applied Measurement in Education*, *1*(4), 279–297.

Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory* (Vol. 1, pp. 109–126). Taylor and Francis.

Nuske, H. J., & Bavin, E. L. (2011). Narrative comprehension in 4-7-year-old children with autism: Testing the Weak Central Coherence account. *International Journal of Language & Communication Disorders*, *46*(1), 108–119.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Pearson, P. D., & Johnson, D. D. (1978). Questions. In *Teaching reading comprehension* (pp. 153–178). Holt, Rinehart and Winston.

Raphael, T. E., & Au, K. H. (2005). QAR: Enhancing comprehension and test taking across grades and content areas. *The Reading Teacher*, *59*(3), 206–221.

Stachl, C. N., & Baranger, A. M. (2020). Sense of belonging within the graduate community of a research-focused STEM department: Quantitative assessment using a visual narrative and item response theory. *PloS one*, *15*(5), e0233431.

Van der Hallen, R., Evers, K., Brewaeys, K., Van den Noortgate, W., & Wagemans, J. (2015). Global processing takes time: A meta-analysis on local-global visual processing in ASD. *Psychological Bulletin*, *141*(3), 549–573.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566-4575).

Von Ahn, L., & Dabbish, L. (2004, April). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 319-326).

Wang, X., Chen, W., Wang, Y. F., & Wang, W. Y. (2018). No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 43–74). Springer.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.