Crowdsourcing Thumbnail Captions via Time-Constrained Methods

CARLOS AGUIRRE, Johns Hopkins University, USA
AMAMA MAHMOOD, Johns Hopkins University, USA
CHIEN-MING HUANG, Johns Hopkins University, USA

Speech interfaces, such as personal assistants and screen readers, employ captions to allow users to consume images; however, there is typically only one caption available per image, which may not be adequate for all settings (e.g., browsing large quantities of images). Longer captions require more time to consume, whereas shorter captions may hinder a user's ability to fully understand the image's content. We explore how to effectively collect both thumbnail captions—succinct image descriptions meant to be consumed quickly—and comprehensive captions, which allow individuals to understand visual content in greater detail. We consider text-based and time-constrained methods to collect descriptions at these two levels of detail, and find that a time-constrained method is most effective for collecting thumbnail captions while preserving caption accuracy. We evaluate our collected captions along three human-rated axes—correctness, fluency, and level of detail—and discuss the potential for model-based metrics to perform automatic evaluation.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; Accessibility design and evaluation methods; Accessibility technologies; • **Information systems** → *Crowdsourcing*; • **Computing methodologies** → *Machine learning*; Computer vision representations; Computer vision tasks;

Additional Key Words and Phrases: image captions, crowdsourcing, accessibility, annotation interfaces

ACM Reference Format:
Carlos Aguirre, Amama Mahmood, and Chien-Ming Huang. 2022. Crowdsourcing Thumbnail Captions via Time-Constrained Methods. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3490099.3511136

1 INTRODUCTION

The rapid expansion of information available online necessitates new avenues of information consumption; for example, browsing through news feeds on social media platforms such as Twitter, Facebook, and Instagram and consuming information via images has become commonplace. "Thumbnail" presentations, which are textual or image previews, enable people to browse through a plethora of information via infinite scroll before deciding what to consume in depth. Ideally, thumbnail presentations provide internet users with just enough information to decide what to consume when navigating the ever-increasing amount of information online.

In contrast, speech interaction users rely on screen readers and image captions when consuming online information. Given that a small percentage of online images have captions available [12], efforts to increase the number of captions online has focused on generating *automatic alternative text* (AAT) [7, 15, 45, 47] and crowdsourcing captions for image datasets [39]. AAT generators and crowdsourcing mechanisms typically generate one caption per image, which does not afford the kind of rapid information browsing that thumbnail images could potentially provide. For instance, Apple devices allow Facebook to announce image alt text for screen reader users [33] and other notifications in its hands-free mode [19]; however, the AAT generated for images is very rudimentary (e.g., "Image may contain tree, sky, outdoor" [33]). Additionally, announcing long or tedious image descriptions may not always be desirable for screen reader users. Hence, thumbnail captions may provide a more useful format for these types of interactions. Another use case of thumbnail captions is the consumption of online information by Blind and Visually Impaired (BVI) individuals; short captions with fewer details, while not enough to fully mentally visualize images, can provide enough flexibility to skim through a plethora of images for information of interest, whereas long, detailed captions may hinder a BVI individual's ability to quickly sift through posts on their timeline.

Recent research has explored methods of interaction with multiple descriptions for one image [24, 30, 31]; in particular, Morris et al. [30] illustrated the potential benefits of multi-level *progressive detail* interaction—wherein the first caption of the progressive detail is meant to be equivalent in detail to standard alternative ("alt") text, while subsequent captions may reveal more information—to support better comprehension and flexibility by providing more control as to how much information a BVI individual wants to receive. Furthermore, it was suggested by the participants in this study that details should be carefully *ordered* to reflect how a sighted person interprets an image in a step-by-step manner. Motivated by how multi-level *progressive detail* image descriptions may be employed to design better interactions for BVI individuals, we explore **how to effectively collect thumbnail captions** to afford faster browsing of imagery for speech interaction users in general.

Providing descriptions with multiple levels of detail for online images is a cumbersome task for alt text authors, especially as it is already unlikely for such image descriptions to be provided in the first place [12, 28]. While crowdsourcing mechanisms come with their own sets of known challenges in obtaining image descriptions [39], previous works have developed effective methods for collecting image descriptions [17, 46], leading to the creation of image caption datasets [7, 15]. However, it is unclear whether *we can leverage similar crowdsourcing mechanisms to obtain thumbnail captions or captions with specific levels of detail.*

As a step toward enabling captions for images with specific levels of information, in this work we are interested in designing crowdsourcing mechanisms to effectively obtain captions with two levels of detail: 1) "thumbnail," short captions containing only the information essential to succinctly describe an image and 2) "comprehensive," long captions describing all aspects of an image in greater detail. To this end, we compare conventional methods for collecting descriptions with online workers, wherein we textually specify the level of detail desired, and a new, time-constrained collection method that limits how much time a worker has to view the image before providing a description. The rationale behind this time-constrained method is that it leverages the most salient information about an image, which humans extract from a quick glance. Specifically, we allow online workers to view an image for only 500 milliseconds, hypothesizing that this timed method will yield thumbnail captions with fewer—yet correct—details. We evaluate the effectiveness of these text-based and time-constrained methods to collect captions with different levels of detail by measuring caption accuracy and detail with both human ratings and model-based metrics. This work makes two main contributions:

- We propose and validate a time-constrained method for the effective collection of thumbnail-style image captions. Thumbnail captions are concise descriptions intended to aid speech interaction users in browsing online images.
- We explore and empirically validate model-based metrics for assessing the human-rated level of detail in image captions with the goal of enabling future automatic evaluations.

2 RELATED WORK

2.1 One Caption Does Not Fit All

Current approaches to image captioning rely on a single image caption—also referred to as alt text or image description—to textually convey information pertaining to an image [45]. However, research involving BVI individuals has highlighted the need for different types of descriptions depending on context and content, demonstrating that one-size-fits-all approaches to image captions are suboptimal for speech interaction users' understanding and consumption of images [13, 18, 20, 24, 28, 30, 41]. While prior research has focused on the delivery of captions via various speech modalities (e.g., varying speed of speech [8]), recent work has explored different methods of transmitting information by providing multiple descriptions per image; for example, Zhong et al. [47] created an interface for region-based image descriptions that stitched together separate images (regions) and their crowdsourcing descriptions. In addition to these "spatial captions," Morris et al. [30] explored novel interactions for consuming images through screen readers, including *progressive detail*; the goal of progressive detail is to give users control over the amount of information and the time they need to understand an image. To achieve this, content authors must provide multiple descriptions for a single image and the logical order in which the screen reader should convey these descriptions; shorter, more essential descriptions would be read first by the screen reader, followed by longer and more comprehensive ones. Morris et al. experimented with three levels of detail and speculated that their implementation was flexible to a different number of levels; additionally, their study participants demonstrated preferences for specific levels of detail (i.e., some participants preferred only two levels). Having two captions available—one short, high-level preview and one long, more comprehensive description—is also suggested by the Web Content Accessibility Guidelines (WCAG) for complex images (e.g., graphs and figures [2]) and other survey studies of BVI individuals' preferences in consuming images [20, 28]. Following these recommendations, recent work has explored creating new speech interactions for image consumption by combining spatial and two-detail-level interactions [24, Image Explorer], as well as evaluation tools for descriptions [18]. In our study, we also consider two levels of detail, *thumbnail* (essential) and *comprehensive*. Rather than focusing on the usefulness of the descriptions to BVI individuals, we first study the effectiveness of the collection methods that can be used at a large scale for crowdsourcing captions with varied levels of detail.

2.2 Image Caption Collection

Creating alt text for images is a challenging task for both content authors and automation. Twitter and other social media services have tried to address accessibility problems by allowing authors of content (posts, stories, and Tweets) to provide alt text to their images; however, only ~0.1% of Tweets with images have alt text, revealing the need for automated methods to collect image descriptions [12].

Prior work has focused on scraping previously existing captions of similar images available online [14, 47] and using human-in-the-loop approaches to produce captions for solicited images [1, 36, 44]. However, these approaches share challenges—including cost, speed, and privacy—suggesting the need for alternate methods [15] such as AAT generators, which are cheaper, faster, and more private than human-in-the-loop collection. Social media platforms such as Facebook have recently begun to deploy AAT generators with positive outcomes [45]. Multiple datasets have been developed to train AAT generators; however, there remains a need to explore methods of collecting such datasets at a larger scale. Microsoft Common Objects in Context (MS COCO) [7], one of the first image caption datasets, set the standard for user interfaces that collect descriptions for images on a large scale through crowdsourcing; since then, most other image caption datasets have used collection methods inspired by the MS COCO crowdsourcing interface [15]. For consistency with prior work, we use a modified version of the MS COCO interface for our *control* collection method. In addition, we explore the effectiveness of collection methods that use either text-based instructions or time-constrained interactions to collect captions with specific levels of detail.

2.3 Image Caption Evaluations

While there is obviously a need for AAT generators, the resulting generated captions are not up-to-par for the BVI community and often require human corrections [37]; furthermore, it has been shown that BVI individuals are overly trusting of automatically generated captions [29], rendering caption quality a critical issue. Quality evaluation is one of the biggest challenges in using large-scale crowdsourcing methods [13]. Prior research focused on the evaluation of image descriptions has mainly been conducted in the setting of collecting datasets for automatic image captioning [7, 11, 16, 21–23, 25]; these methods utilize a wide range of natural language processing tools, from relatively simple n-gram matching algorithms like bilingual language evaluation understudy (BLEU) [32] to vision-language trained models such as ViLBERTScore [23]. However, the majority of these methods of evaluating image captions rely on matching them against reference descriptions [11, 21].

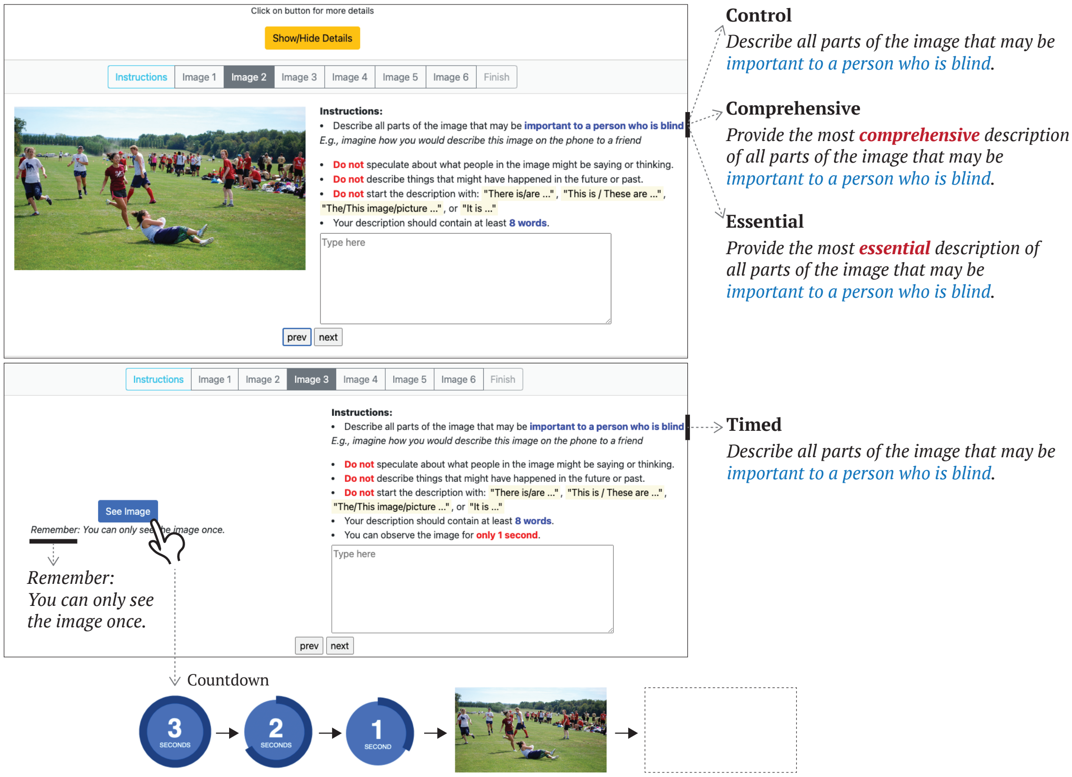
Due to the subjective nature of image captions (i.e., there may be several correct descriptions for a single image), it is standard practice to include more than one correct caption when evaluating image captioning datasets (*reference* caption). For instance, MS COCO [7] collected five reference captions per image and employed commonly used metrics for text similarity to evaluate the distance (similarity) between the candidate and reference captions [7]. However, these initial metrics, including BLEU [32], METEOR [10], and ROUGE [26], are sub-optimal for comparing similar (rather than identical) descriptions, as they suffer from common pitfalls such as not recognizing synonyms and over-sensitivity to overlapping n-grams of otherwise semantically distinct descriptions [3]; these limitations have also been observed in machine translation and sister tasks, where these metrics were first employed. Since the development of these metrics, there has been considerable progress on new metrics for text similarity and specifically evaluating image captions [4, 16, 22, 23, 25] which attempt to address these issues, although the original metrics remain the most commonly used. In this study, we expand on how some of these new metrics perform when compared to human evaluations.

3 METHODS

3.1 Collection Methods for Obtaining Image Descriptions

In this study, we explore text-based and time-constrained methods for collecting image descriptions. Fig. 1 depicts our online interface used to collect captions via the following four methods that we experimentally studied in this work:

- **Control.** The *control* collection method uses the instructions from MS COCO [7] with three modifications to adapt the task for BVI individuals: 1) the inclusion of "important to a person who is blind" [15] in the main instruction; 2) the removal of the instruction that prohibits proper names in descriptions, as BVI individuals prefer descriptions with proper names if an image contains a famous person [40]; and 3) the removal of the instruction that limits captions to at most one sentence to allow for greater flexibility in providing different levels of detail.
- **Text-based comprehensive information** (Comprehensive). The *comprehensive* collection method adds to the basic instructions by prompting workers to write "the most *comprehensive*" description of the image. We expect descriptions collected via this method to be longer in length and contain more detail as compared to those collected via the control method.



Instructions:

- Describe all parts of the image that may be **important to a person who is blind**

E.g., imagine how you would describe this image on the phone to a friend

- **Do not** speculate about what people in the image might be saying or thinking.
- **Do not** describe things that might have happened in the future or past.
- **Do not** start the description with: "There is/are ..." , "This is / These are ..." , "The/This image/picture ..." , or "It is ..."
- Your description should contain at least **8 words**.

Instructions:

- Describe all parts of the image that may be **important to a person who is blind**

E.g., imagine how you would describe this image on the phone to a friend

- **Do not** speculate about what people in the image might be saying or thinking.
- **Do not** describe things that might have happened in the future or past.
- **Do not** start the description with: "There is/are ..." , "This is / These are ..." , "The/This image/picture ..." , or "It is ..."
- Your description should contain at least **8 words**.
- You can observe the image for **only 1 second**.

Remember: You can only see the image once.

Remember:
You can only see
the image once.

Control

Describe all parts of the image that may be important to a person who is blind.

Comprehensive

Provide the most ***comprehensive*** *description of all parts of the image that may be important to a person who is blind.*

Essential

Provide the most ***essential*** *description of all parts of the image that may be important to a person who is blind.*

Timed

Describe all parts of the image that may be important to a person who is blind.

Fig. 1. Our online interface for obtaining image captions via four collection methods: control, comprehensive, essential, and timed. For the first three collection methods (control, comprehensive and essential), the text instructions specify the required detail level of the caption (top). For the timed collection method (bottom), when the user selects the *See Image* button, the image appears following a countdown of 3 seconds and disappears after a fixed interval of 500 milliseconds.

- **Text-based essential information** (Essential). The *essential* collection method adds to the basic instructions by prompting workers to write "the most *essential*" description of the image. We expect these descriptions to be shorter in length and contain fewer details as compared to the control descriptions.
- **Time-constrained** (Timed). The *timed* collection method has the same instructions as the control method but limits the time the workers can view an image to 500 milliseconds, analogous to how a sighted person would briefly view an image while browsing online. To decide the length of the time constraint, we referred to prior work on human signal processing, which has found that humans are able to *classify* images in under 150 milliseconds [42]. We ran small pilot studies amongst colleagues to test task difficulty for 500-, 1,000-, 3,000-, and 5,000-millisecond time windows; with no clear variance in task difficulty reported by participants, we chose the smallest time window, 500 milliseconds.

3.2 Experimental Task and Procedure

We collected image captions from online workers on Amazon Mechanical Turk as following prior research [7, 15]. For each collection method, workers were instructed to provide descriptions for sets of images from the MS COCO captions dataset [7]. As recommended by prior work [40], we created four sets of six images selected by the authors to balance

Questions	Examples
Is the description gibberish?	re re yry yh rty defgfdhj th tryhtr5 er yeg
Is the description clearly incorrect English?	There was a people is bus vacation travelling.
Are there repeated variations of the same description in the assignment?	no color blind see the picture in the image I am see the picture in the image of no color blind Good picture in the see the man is no color blind Picture see the image in the man no color blind See the picture in the image in the man of no blind Picture see the image of the man no color blind
Is the description an opinion?	I like the image of the man.

Table 1. The exclusion criteria we used to remove clearly low-quality descriptions obtained from Amazon Mechanical Turk workers.

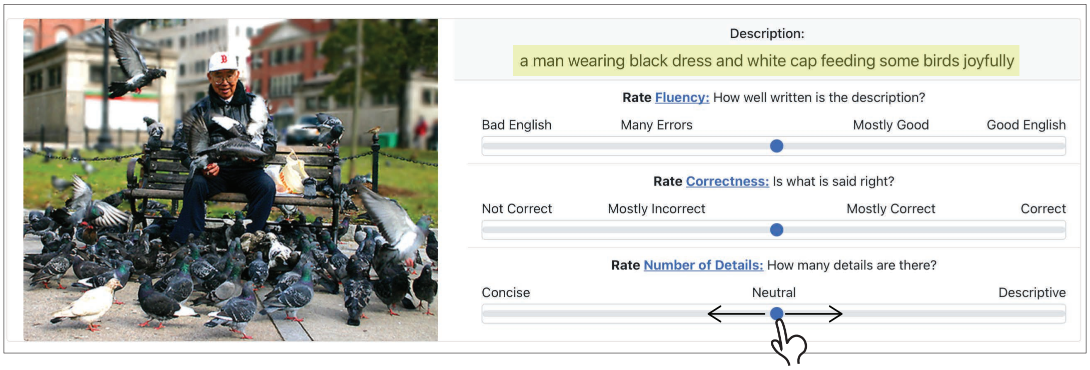
key themes in images commonly found online: event/scene, people, and objects/landmarks. Each of the four sets—four Human Intelligence Tasks (HIT)—contained two images per theme ($2 \times 3 = 6$ images in each set), resulting in a total of 24 (6×4) distinct images for this study. While a single worker was allowed to provide captions using all four collection methods, our implementation restricted workers from repeating collection methods or image sets; each image set was randomly assigned to one collection method for a worker who chose to do more than one HIT. However, we observed that the majority of the workers provided image captions using only one of the four collection methods (one HIT).

3.3 Evaluation Metrics

While description quality is one of the biggest challenges in using crowdsourcing for obtaining image captions at scale [13], in our setting, we had the additional challenge of measuring the two levels of description detail. Furthermore, previous works recommend correlating automatic evaluations with human judgments in new settings [16, 23]; therefore, we considered both model-based metrics and human evaluations for quantifying image quality (as measured by correctness and fluency) and the amount of detail provided in the image descriptions.

3.3.1 Exclusion criteria. To ensure the basic quality of our collected image descriptions, we manually checked for clearly inappropriate descriptions, such as those including random characters, repeated descriptions, and those with clearly poor English, as determined by the authors. Table 1 lists our exclusion criteria along with examples of inferior descriptions.

3.3.2 Caption consistency. One of the challenges in obtaining image captions via crowdsourcing is that descriptions are subjective; multiple workers may produce different, but perhaps equally correct, captions. Furthermore, captions at different levels of detail may display different degrees of similarity based on the level of detail each individual worker chooses to include; for instance, when asked to provide a succinct description of an image with multiple objects of interest, workers may choose to include different subsets of objects, resulting in artificially lower caption similarity scores. Therefore, we first quantified the similarity of descriptions collected from different online workers for each image and collection method. Recently, new *model-based* metrics have been introduced, which may reduce concerns from earlier works on caption similarity as discussed in Section 2.3. Reimers and Gurevych [35] developed a sentence-level BERT (Sentence-BERT) Siamese model that finds the cosine distance between the vector representations of two sentences and has been shown to outperform other embedding-based methods. In our study, we used Sentence-BERT to compute the average overall pairwise cosine similarity between descriptions, where a score of 1 signifies that the



Description:

a man wearing black dress and white cap feeding some birds joyfully

Rate Fluency: How well written is the description?

Rate Correctness: Is what is said right?

Rate Number of Details: How many details are there?

Fig. 2. Our interface for human judges to rate fluency, correctness, and amount of detail in the collected captions.

descriptions are the same and 0 indicates that the vectors are completely perpendicular (i.e., the descriptions are very different).

3.3.3 Human evaluations. Prior research on human evaluation of automatically generated texts has explored manual ratings across multiple evaluation dimensions, such as correctness, informativeness, and fluency [25, 43]. We adapted a similar approach by asking human judges to rate a subset of the online workers' descriptions across three dimensions: fluency, correctness, and amount of detail. Such human evaluation not only allows us to evaluate the effectiveness of each collection method, but also various model-based metrics for quantifying correctness and the amount of detail in the obtained captions. Fig. 2 illustrates the user interface we used to collect human evaluation ratings.

- **Fluency** (0–100). The fluency of the language in a description is an important aspect of its quality, as systematic language errors (e.g., misspellings, grammatical errors) add further challenges to the training of AAT generators. Additionally, for this study, language errors such as misspellings could skew our language model-based metrics. To assess fluency, judges were instructed to rate a description's language fluency on a scale of 0 to 100 and focus on the text only, ignoring the image. They were also instructed to take into account spelling errors, bad grammar, and awkward word usage in their ratings.
- **Correctness** (0–100). The correctness (or appropriateness) of the details in a description is another aspect of image description quality; to assess correctness, judges were instructed to focus on both the description and its corresponding image and rate whether the details contained in the description were correct. However, they were specifically advised not to focus on the length of or amount of detail in the description.
- **Amount of detail** (0–100). To evaluate the effectiveness of various collection methods in obtaining two levels of detail (*essential* and *comprehensive*), we asked judges to focus on the textual description alone and rate the amount of detail contained in the description regardless of its correctness.

3.3.4 Model-based metrics. As the size of image description datasets continues to increase, the need for automatic quality assessment tools becomes vitally important. Below, we describe two categories of metrics—correctness and amount of detail—and the specific model-based metrics we employed in this work.

Correctness metrics. We utilized two model-based metrics for caption quality estimation, both of which have been previously shown to be well-correlated with human judgments of correctness.

- **SPICE_f** (0–1). Simple metrics for image captions consider candidate descriptions based on n-gram overlaps with reference descriptions; however, these metrics suffer from common pitfalls, as discussed in Section 2.3. SPICE_f was created to target these limitations by using dependency trees, which abstract away most lexical and syntactic idiosyncrasies, and then employing a rule-based method to map these trees into "scene" graphs to make semantically-rich candidate-reference comparisons [3]. SPICE_f is an "F" measure, ranging from 0.0 (for incorrect captions) to 1.0 (for correct captions). The previously reported correctness cut-off point is approximately .19 for the MS COCO dataset [3]. One of the limitations of this metric is that it is highly dependent on both the quality of the dependency parser used to create the trees and the generalizability of the rules manufactured by the authors of the model. Despite this, SPICE_f has been shown to be well-correlated with human judgments on various domains and datasets [16, 21].
- **ViLBERTScore_f** (0–1). SPICE_f is based on reference descriptions only and not on images themselves; therefore, it is subject to failure when the available reference descriptions are of poor quality. Contrarily, ViLBERTScore_f [23] considers both caption and image to generate a correctness measure. Using the joint vector representations of the image and its captions from a pre-trained ViLBERT model [27], ViLBERTScore_f [23] computes the cosine similarity of the vector representations for the reference caption + image and the candidate caption + image to obtain a metric between 0–1, where 1.0 is a correct caption and 0.0 is an incorrect caption. ViLBERTScore_f has been found to be significantly more correlated with human judgment on multiple benchmark datasets than all previous metrics [23]; however, it is still dependent on reference captions, rendering it vulnerable to poor-quality text references. Additional methods that do not require reference descriptions are still very new, and at the time of writing, their code and pre-trained models are not yet publicly available [16, 22]; however, initially reported results suggest that ViLBERTscore_f remains competitive.

Detail level metrics. Collecting captions at different levels of detail is a novel task and the best method of measuring such levels of detail has not yet been determined. In this work, we use two model-based metrics to approximate the amount of detail contained in a description: the number of words in noun phrases (NPs) and cross-entropy.

- **Number of words in NPs.** NPs are groups of words that contain a noun and its modifiers. "Number of words in NPs" is a reflective measure of how many details are packed into a description. Although prior work has introduced metrics that utilize multiple parts-of-speech tags, it is still unclear which parts of speech are the most important for measuring detail level. Moreover, using such amalgamated metrics creates multiple axes of comparison, turning it into an even more convoluted problem; in contrast, counting the words in NPs is a single-axis metric with a clear indication of detail. Thus, we used an NP tagger from the Natural Language ToolKit (NLTK) [5] to count the words contained in each NP. While most words in NPs are typically nouns and thus important for measuring detail level, determiners in NPs (such as "the") are informative but common across different types and styles of descriptions; given that they are equally frequent, we assume that determiners do not affect relative comparisons disproportionately.
- **Cross-entropy.** Cross-entropy is a widely used metric in the NLP community, employed to measure the information conveyed in a sentence and to optimize language models toward a reference distribution ever since Shannon's introduction of the metric [38]. Cross-entropy measures the distance of observed events and a language distribution where highly likely events yield low information (e.g., the word "the") and unlikely events yield high information (e.g., the word "platypus"). The cross-entropy of a description is calculated by adding the log probability of the words in a description given its context. While accounting for word usage, this metric correlates with the length of a description; for example, the phrase "the table" (p_1) will have a lower entropy

than "the brown table" (p_2) because p_2 conveys more detail than p_1 by including the word "brown" to describe the table's color. Similarly, p_2 will have a lower entropy compared to "the fuschia table" (p_3) because that color is less likely to occur in text describing a table. We computed the cross-entropy of an image description using GPT-2 Large [34] to measure the information conveyed based on the learned language distribution of GPT-2;[1] we hypothesize that more complex descriptions (higher entropy) contain more detail. The goal of this metric (along with using the number of words in NPs) is to estimate detail in order to differentiate between *essential* and *comprehensive* image descriptions. A limitation of using cross-entropy is that it is highly dependent on the language distribution learned by the model, which in turn depends on its implementation and, more importantly, the data that it was trained on. However, GPT-2 Large was trained on a fairly large corpus from the Internet containing a varied range of topics, thus theoretically making it less prone to biases based on training data domain and/or observed levels of detail.

4 RESULTS

For the results reported below, we used a one-way analysis of variance (ANOVA) where the collection method was set as the fixed effect. All post hoc pairwise comparisons were conducted using independent t-tests with Bonferroni corrections. For all statistical tests reported below, $p < .05$ is considered a significant effect; we followed Cohen's guidelines on effect size and considered $\eta_p^2 = 0.01$ a small effect size, $\eta_p^2 = 0.06$ a medium effect size, and $\eta_p^2 = 0.14$ a large effect size [9].

We recruited a total of 69 online workers to generate image descriptions and collected a total of 768 captions. Examples of collected descriptions and their scores from our human judges and model-based metrics are shown in Fig. 3. The majority of workers performed only one HIT, or six descriptions, with an average of 11 descriptions overall. Since our algorithm assigned collection methods randomly, it was unlikely to obtain the same number of descriptions for each combination of image set and collection method. We grouped the descriptions by image set and collection method in Fig. 4.

Caption consistency. To verify that workers produced comparably consistent captions across the collection methods, we compute cosine similarity using Sentence-BERT as a variance metric for captions grouped by image. A one-way ANOVA reveals no significant main effect of the collection method on the metric of cosine similarity, $F(3, 92) = 0.17, p = .916, \eta_p^2 = .006$ (Fig. 5, left).

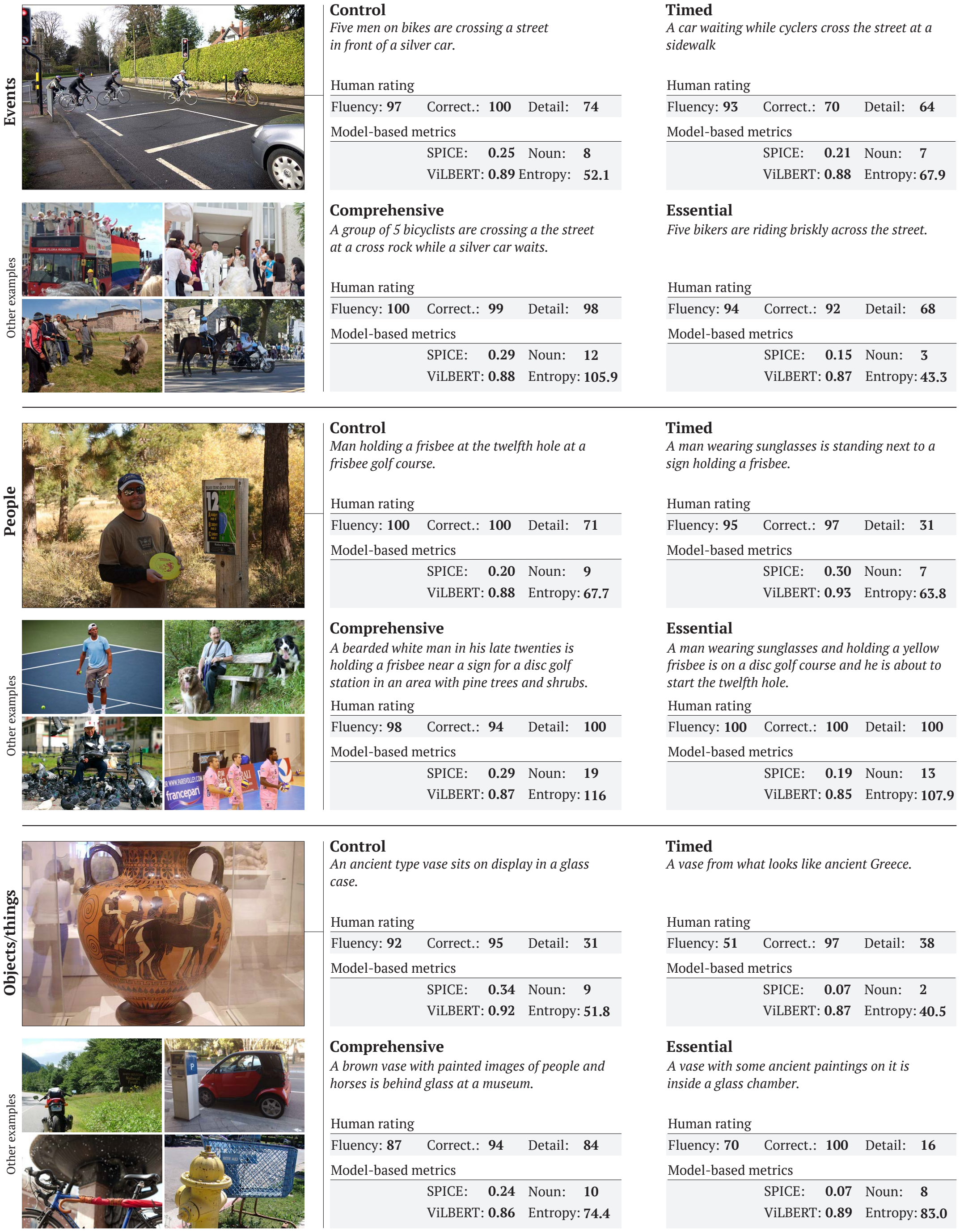
4.1 Human Evaluation

We obtained human evaluations for a random set of 360 descriptions collected via various annotation methods. A total of 11 (5 female, 5 male, 1 undisclosed) human judges aged 20 to 25 ($M = 22.9, SD = 1.66$) were recruited via reading and laboratory group meetings from a pool of both undergraduate and graduate students at the authors' institution.[2] All judges self-reported their English proficiency as professional working or better; a majority (70%) reported native English proficiency. Fig. 5 visualizes the main results of these human evaluations.

A one-way ANOVA revealed no significant main effect of the annotation method on the rating of caption fluency, $F(3, 356) = 0.30, p = .828, \eta_p^2 = .002$. Additionally, a one-way ANOVA indicated no significant main effect of annotation method on caption correctness, $F(3, 356) = 1.02, p = .38, \eta_p^2 = .009$. However, we observed a significant main effect of the annotation method on the amount of detail in the collected captions, $F(3, 356) = 5.91, p < .001, \eta_p^2 = .047$.

[1] GPT-2 (and Large) are transformer-based language models typically used for text generation [34]

[2] Note that one judge chose not to share their demographics.



Control
Five men on bikes are crossing a street in front of a silver car.

Human rating
Fluency: **97** Correct.: **100** Detail: **74**

Model-based metrics
SPICE: **0.25** Noun: **8**
ViLBERT: **0.89** Entropy: **52.1**

Timed
A car waiting while cyclers cross the street at a sidewalk

Human rating
Fluency: **93** Correct.: **70** Detail: **64**

Model-based metrics
SPICE: **0.21** Noun: **7**
ViLBERT: **0.88** Entropy: **67.9**

Comprehensive
A group of 5 bicyclists are crossing a the street at a cross rock while a silver car waits.

Human rating
Fluency: **100** Correct.: **99** Detail: **98**

Model-based metrics
SPICE: **0.29** Noun: **12**
ViLBERT: **0.88** Entropy: **105.9**

Essential
Five bikers are riding briskly across the street.

Human rating
Fluency: **94** Correct.: **92** Detail: **68**

Model-based metrics
SPICE: **0.15** Noun: **3**
ViLBERT: **0.87** Entropy: **43.3**

Control
Man holding a frisbee at the twelfth hole at a frisbee golf course.

Human rating
Fluency: **100** Correct.: **100** Detail: **71**

Model-based metrics
SPICE: **0.20** Noun: **9**
ViLBERT: **0.88** Entropy: **67.7**

Timed
A man wearing sunglasses is standing next to a sign holding a frisbee.

Human rating
Fluency: **95** Correct.: **97** Detail: **31**

Model-based metrics
SPICE: **0.30** Noun: **7**
ViLBERT: **0.93** Entropy: **63.8**

Comprehensive
A bearded white man in his late twenties is holding a frisbee near a sign for a disc golf station in an area with pine trees and shrubs.

Human rating
Fluency: **98** Correct.: **94** Detail: **100**

Model-based metrics
SPICE: **0.29** Noun: **19**
ViLBERT: **0.87** Entropy: **116**

Essential
A man wearing sunglasses and holding a yellow frisbee is on a disc golf course and he is about to start the twelfth hole.

Human rating
Fluency: **100** Correct.: **100** Detail: **100**

Model-based metrics
SPICE: **0.19** Noun: **13**
ViLBERT: **0.85** Entropy: **107.9**

Control
An ancient type vase sits on display in a glass case.

Human rating
Fluency: **92** Correct.: **95** Detail: **31**

Model-based metrics
SPICE: **0.34** Noun: **9**
ViLBERT: **0.92** Entropy: **51.8**

Timed
A vase from what looks like ancient Greece.

Human rating
Fluency: **51** Correct.: **97** Detail: **38**

Model-based metrics
SPICE: **0.07** Noun: **2**
ViLBERT: **0.87** Entropy: **40.5**

Comprehensive
A brown vase with painted images of people and horses is behind glass at a museum.

Human rating
Fluency: **87** Correct.: **94** Detail: **84**

Model-based metrics
SPICE: **0.24** Noun: **10**
ViLBERT: **0.86** Entropy: **74.4**

Essential
A vase with some ancient paintings on it is inside a glass chamber.

Human rating
Fluency: **70** Correct.: **100** Detail: **16**

Model-based metrics
SPICE: **0.07** Noun: **8**
ViLBERT: **0.89** Entropy: **83.0**

Fig. 3. Examples of worker captions obtained using the four collection methods alongside their human ratings and model-based metrics .

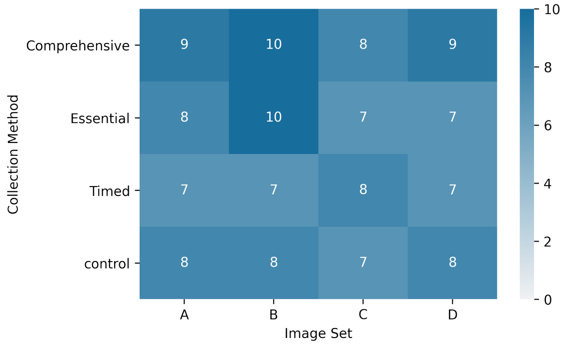


Fig. 4. Number of descriptions obtained per image set, arranged by collection method. Images are grouped in four sets (A, B, C, D) of 6 images each. We obtained a minimum of 7 candidate descriptions per image set for each collection method.

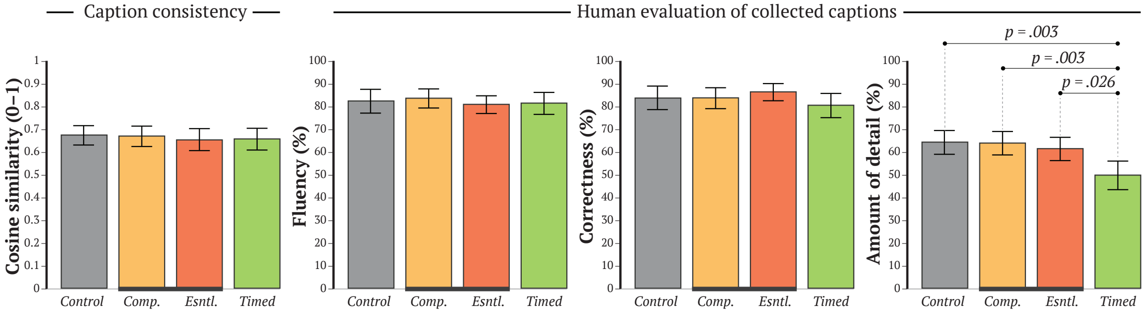


Fig. 5. Results of caption consistency (left) and human evaluation (right). One-way ANOVAs were conducted to discover the effects of collection method on caption consistency (cosine similarity) and on human ratings of fluency, correctness, and amount of detail. Error bars represent 95% confidence intervals; only significant comparisons ($p < .05$) are highlighted.

Pairwise independent t-tests with Bonferroni corrections revealed that participants rated the *timed* method's captions ($M = 49.9, SD = 28.37$) as less detailed compared to all other methods': *control* ($M = 64.4, SD = 23.26$), $p = .003$, *comprehensive* ($M = 64.0, SD = 26.26$), $p = .003$, and *essential* ($M = 61.5, SD = 24.59$), $p = .026$.

4.2 Model-Based Evaluation

Fig. 6 visualizes the main results of our model-based metrics.

Correctness. To computationally quantify the correctness of the descriptions collected by different annotation methods, we used the ViLBERTScore_F and SPICE_F metrics. A one-way ANOVA revealed no significant main effect of the annotation method on ViLBERTScore_F, $F(3, 764) = 1.84, p = .138, \eta_p^2 = .007$; we also did not observe a significant main effect of the annotation method on SPICE_F, $F(3, 764) = 2.12, p = .097, \eta_p^2 = .008$.

Amount of detail. To automatically gauge the amount of detail in the collected descriptions, we used two metrics: number of words in NPs and cross-entropy. A one-way ANOVA revealed a significant main effect of the annotation method on the number of words in NPs, $F(3, 764) = 6.47, p < .001, \eta_p^2 = .025$. A pairwise independent t-test with Bonferroni corrections showed that participants rated the *timed* method ($M = 7.49, SD = 2.34$) as resulting in fewer words in NPs as compared to all other methods: *control* ($M = 8.56, SD = 3.99$), $p = .013$, *comprehensive* ($M = 8.74, SD = 3.41$),

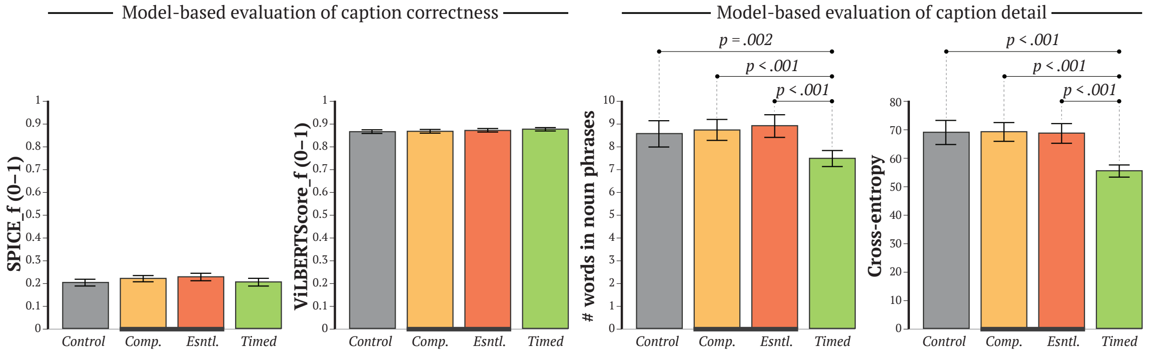


Fig. 6. Results of model-based evaluations of caption correctness and detail. One-way ANOVAs were conducted to discover the effects of collection method on caption correctness (SPICE_f and ViLBERTScore_f) and detail (number of words in NPs and cross-entropy) as evaluated via model-based metrics. Error bars represent 95% confidence intervals; only significant comparisons ($p < .05$) are highlighted.

$p < .001$, and *essential* ($M = 8.91, SD = 3.46$), $p < .001$. Additionally, a one-way ANOVA revealed a significant main effect of the annotation method on cross-entropy, $F(3, 764) = 14.45, p < .001, \eta_p^2 = .054$. A pairwise independent t-test with Bonferroni corrections showed that participants rated the *timed* method ($M = 55.5, SD = 13.87$) as having lower cross-entropy than all other methods: *control* ($M = 69.0, SD = 29.14$), $p < .001$, *comprehensive* ($M = 69.2, SD = 24.44$), $p < .001$, and *essential* ($M = 68.7, SD = 24.09$), $p < .001$.

4.3 Metric Correlations

To validate the results of our model-based metrics, we compared them to our human evaluation results, as suggested by prior research [3, 16, 23]. Fig. 7 summarizes the results of our correlation analyses.

Correctness. We computed Pearson (r) correlations to assess the relationships between the *correctness* human rating and our model-based metrics for correctness: ViLBERTScore_f and SPICE_f. There was a positive correlation between the human rating of correctness and ViLBERTScore_f ($r(358) = .18, p < .001$) and between the human rating of correctness and SPICE_f ($r(358) = .29, p < .001$). Additionally, we computed the Pearson correlation to assess the relationship between our model-based metrics; there was a positive correlation between SPICE_f and ViLBERTScore_f ($r(358) = .40, p < .001$).

Amount of detail. We also computed Pearson (r) correlations to assess the relationships between the human rating of *detail* and our corresponding model-based metrics. There was a positive correlation between the human rating of caption detail and the number of words in NPs ($r(358) = .48, p < .001$) and between the human rating of caption detail and cross-entropy ($r(358) = .48, p < .001$). Finally, we computed the Pearson correlation between our model-based metrics to measure detail level; there was a positive correlation between cross-entropy and number of words in NPs ($r(358) = .66, p < .001$).

4.4 Additional Exploration

In this section, we report the results of exploratory analyses motivated by the results reported above.

Time spent writing descriptions. The time-constrained collection method limited the amount of time that workers were allowed to observe the image; however, it did not limit the amount of time workers were allowed to write their

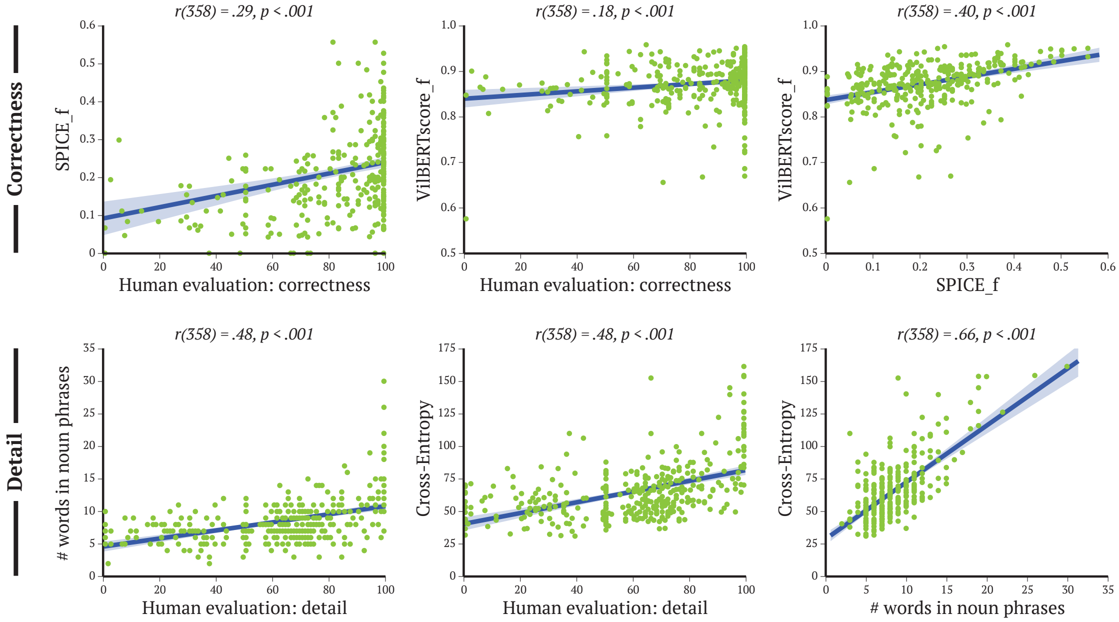


Fig. 7. Scatter plots illustrating the correlation between model-based metrics (SPICE_f and ViLBERTScore_f) and human ratings for correctness (top), between model-based metrics (number of words in NPs and cross-entropy) and human ratings for detail (bottom), and amongst model-based metrics for correctness (SPICE_f vs. ViLBERTScore_f) and detail (cross-entropy vs. number of words in NPs) (right).

captions. In fact, none of the collection methods limited the amount of time that workers could take to write any single description[3]. We explored whether the collection method influenced the amount of time that workers took to write their captions. Time was measured, in seconds, as the amount of time that workers had the description text box selected; this way, we were not measuring the amount of time that they took to view an image before beginning to write its description. A one-way ANOVA revealed a significant main effect of the collection method on the time spent writing descriptions, $F(3, 764) = 7.53, p < .001, \eta_p^2 = .029$. Pairwise independent t-tests with Bonferroni corrections showed that workers with the *timed* collection method ($M = 36.9, SD = 20.70$) spent less time writing descriptions than compared to all other methods: *control* ($M = 47.7, SD = 29.76$), $p < .001$, *comprehensive* ($M = 51.2, SD = 36.49$), $p < .001$, and *essential* ($M = 46.7, SD = 30.23$), $p = .002$.

Correlation between description length and correctness metrics. While we wanted to collect image descriptions at multiple levels of detail, we also endeavored to achieve comparable correctness for both thumbnail and comprehensive captions. However, the model-based metrics we utilized have only been evaluated with single-detail-level captions previously; moreover, these metrics make use of reference descriptions, which also have a single level of detail. Therefore, it is still unknown whether these metrics are consistent across captions of multiple levels of detail. We explored the relationship between description length, as measured by number of words, and our correctness metrics. A Pearson correlation analysis assessing the relationship between the human correctness rating and caption length found no significant correlation, $r(358) = .08, p = .124$; a Pearson correlation analysis assessing the relationship

[3] Workers had a maximum of 12 minutes to finish a HIT consisting of 6 descriptions.

between SPICE_f and caption length also indicated no correlation, $r(358) = .06, p = .271$. However, we found a negative correlation between ViLBERTScore_f and caption length, $r(358) = -.49, p < .001$.

5 DISCUSSION

5.1 Obtaining Thumbnail Captions with Time Constraints

Our main research question asks whether we can collect captions with specific levels of detail for image datasets through crowdsourcing; in our work, we consider two levels of detail: thumbnail (essential) and comprehensive. Our results reveal that our time-constrained (*timed*) collection method was most effective at collecting image descriptions with fewer details while still maintaining correctness and language fluency (as compared to the other three collection methods). We initially predicted that the timed method might result in erroneous captions, as limited image observation time may hinder an online worker's ability to fully process an image; however, we observed no statistical difference in the fluency and correctness of the captions collected via this method as assessed by both human ratings and model-based metrics. Though our *timed* method did not instruct workers to produce captions at a specific level of detail, the direct limitation of observation time appeared to effectively leverage their natural visual processing ability to extract the most essential information from an image; accordingly, our study suggests that the *timed* collection method is effective at collecting image descriptions for thumbnail-style captions.

From the examples of collected captions shown in Fig. 3, we see that, in most cases, the *timed* descriptions are shorter than those produced by the other methods; however, this is not always the case and caption length does not always accurately indicate level of detail. For example, in the *Events* image example, we see that even though the *timed* description appears longer than the *essential* description, the *timed* caption omits some specific details (e.g., the number of cyclists); similarly, in the *People* example, the *timed* caption does not include the same detail about the sign indicating the twelfth hole of the disc golf course. Furthermore, other factors, such as workers' personal writing styles, may influence the length of written descriptions. These examples suggest that the relationship between a caption's length and its level of detail is complex. Despite longer captions in some cases, the amount of information contained in the time-restricted captions was on average less than other methods while still maintaining similar correctness, which may be explained by the exclusion of more specific details.

We originally conjectured that captions collected via the *timed* method might vary substantially from worker to worker, as different individuals may focus on different aspects of images during the limited observation window. However, our results show that the consistency across captions generated via the *timed* method, measured by the cosine similarity of the vector representations of the captions, was comparable to the similarity of the captions collected with the other methods; this implies that workers may have implicitly focused on the same parts of the images or that they all recollected similar details when producing captions, providing further evidence that time-constrained collection methods may be useful in obtaining consistent captions across a large number of workers.

Finally, the *timed* collection method was more efficient than the other methods based on the time workers spent writing descriptions (see Section 4.4). Their efficient behavior may be a byproduct of the limitation on observation time, as workers may have written the image description more quickly when forced to use memory recall and without the opportunity to check the image again and again; in contrast, workers using the other collection methods had unlimited access to the image while writing their captions. Overall, the time-constrained method is effective for collecting quality thumbnail captions while supporting the efficient acquisition of correct image descriptions.

5.2 Inadequacy of Text-Based Instructions in Acquiring Captions with Varying Details

We originally hypothesized that the text-based *essential* method could also effectively collect thumbnail descriptions; however, our results reveal that the *essential* collection method yielded descriptions with similar levels of detail to those obtained via the *control* method. Similarly, despite the added prompt asking for more detailed descriptions in the text-based instructions, the *comprehensive* collection method *also* yielded captions with detail levels similar to those produced by the *control* method. These findings suggest that the common text-based collection methods compared in our study fail to elicit specific levels of detail from online workers.

To ensure that the workers paid attention to all of the task instructions, we required them to read and check off individual instructions via checkboxes before proceeding to the task. The majority of the workers completed only one HIT (using one collection method to write captions for a set of six images), which makes it less likely that our results may be attributed to the inattention of workers who had, for example, skimmed through the instructions assuming that they were the same as those for a different collection method used previously.

Text-based instructions are also vulnerable to potential differences in language interpretation; the terms "essential" and "comprehensive" could have been understood differently than we had originally intended. For example, the term "essential" could have been misinterpreted as an indication to include *all* important details, which would yield longer descriptions than desired; some examples of this possibility can be seen in Fig. 3, where the *essential* descriptions for the example images of *People* and *Objects/things* are longer than those produced by the control method. Moreover, current crowdsourcing mechanisms do not provide feedback to workers in real time to inform them if their responses are sufficiently "essential" or "comprehensive," as also previously observed by Simons et al. [39]. Overall, text-based instructions had a limited impact on workers' task completion process as compared to the time-constrained method; asking workers to provide appropriate details requires more effort than simply changing the wording of instructions, as humans are prone to revert to their own natural writing tendencies.

5.3 Validation of Model-Based Metrics for Assessing Caption Correctness and Detail Level

The gold standard of validating image descriptions is through human ratings; however, they are costly to obtain compared to automatic evaluations. Furthermore, automatic evaluations may be used in many scenarios: to provide instant feedback to crowdworkers, solving one of the challenges of crowdsourcing image captions [39], or to aid in the training of new AAT generators, as research suggests that contemporary AAT generators fail to satisfy BVI individuals' needs [12] and are not designed to produce captions at different levels of detail (e.g., [45]). However, automatic evaluations are imperfect, often suffering from multiple sources of bias [6], and thus first must be validated against human ratings.

We validated our model-based metrics against their corresponding human ratings. Overall, there was a moderate positive correlation between the human rating of detail and both metrics of *number of words in NPs* and *cross-entropy* (Fig. 7). These two metrics were also positively correlated themselves; this strong correlation is expected, as both metrics are directly related to description length—a longer string of text typically has a lower probability (higher cross-entropy) in a language model due to the combinatorics of words, and is also more likely to have more words in its noun phrases. These correlative findings suggest that both the number of words in NPs and cross-entropy are adequate automatic metrics for the evaluation of detail level in descriptions.

However, there was a weak positive correlation between the human rating of correctness and both SPICE_f and ViLBERTScore_f; this finding is in contrast with prior work, where both SPICE_f and ViLBERTScore_f have been shown to have a moderate to strong correlation with human ratings [3, 16, 23]. As described in Section 3.1, a limitation of both of these metrics is that they utilize reference captions, which we collected with the original MS COCO captions dataset; hence, these metrics could be biased toward the level of detail in the reference captions. To explore this possible bias, we calculated the correlation coefficient of both ViLBERTScore_f and SPICE_f against caption length in Section 4.4; we found that ViLBERTScore_f had a moderate negative correlation with description length, while SPICE_f and human correctness had no such significant correlation. Altogether, we found evidence indicating that ViLBERTScore_f is biased against longer captions; this finding suggests the need for new model-based metrics for quantifying correctness that are better correlated with human judgements at different levels of detail. One possible solve is to use metrics that do not depend on reference captions, such as CLIPScore [16], which would theoretically eliminate the problem of reference caption bias. However, these model-based metrics are still dependent on training data, since models that do not explicitly include captions at different levels of detail could still be at risk of detail level bias.

5.4 Limitations and Future Work

In this work, we focused on evaluating the correctness of and level of detail included in an image description; however, we did not measure the *usefulness* of such descriptions in understanding their associated images in specific contexts. As one of the use cases of thumbnail captions is enabling the faster browsing of visual information for screen reader users, an evaluation involving BVI individuals directly must be performed before implementing our methods for data collection in further efforts targeted at that community. While prior work has demonstrated that *progressive detail* interactions are effective for this population [30], it has yet to be determined whether time-constrained caption collection methods are effective at capturing thumbnails that satisfy the needs of BVI individuals. Future studies can utilize a combination of surveys and interviews [24, 28, 30] to evaluate image captions gathered for BVI individuals via various crowdsourcing methods.

Additionally, we limited our study of a time-constrained collection method to 500 milliseconds of image observation time—yet this may not be the most effective time constraint for the collection of thumbnail captions; while we explored longer time constraints in our pilot studies, we did not test a *shorter* time limit. Potentially superior time constraint limits might be inferred from prior work, where workers were given 150 milliseconds to complete image classification tasks [42]. Future work may also explore the use of gaze tracking to better understand how people process visual information in limited periods of time and its relationship to the amount of detail provided in a caption.

Finally, while we were successful in collecting thumbnail captions, we had limited success in collecting more extensive captions with levels of detail beyond those generated via the control method. We speculate that the text prompts employed in this study may not be effective in eliciting detailed captions; further experimentation with the wording of said text prompts is required to gather captions with different levels of detail. Prior work has illustrated that access to captions with multiple levels of detail is beneficial to BVI individuals [20, 28, 30]; future work may explore new methods and mechanisms that support the effective collection of captions at additional levels of detail—including both less and more detail than what is found in typical alt text.

6 CONCLUSION

We explored the collection of image descriptions using time-constrained and text-based methods for two levels of detail: "thumbnail" and comprehensive captions. Our results demonstrate that our time-constrained method is effective at collecting thumbnail descriptions while maintaining comparable correctness and fluency; moreover, we found the text-based methods to be ineffective at collecting descriptions at multiple levels of detail. We validated several model-based metrics for detail level and correctness to enable automatic evaluations, which may be useful for immediate crowdsourcing feedback and quality control in large datasets; for caption correctness, we observed that the model-based metrics had a weak correlation with human ratings due to differences in description lengths at multiple levels of detail. Finally, our study highlights that both of our model-based metrics of detail level have a moderate positive correlation with human ratings, validating these metrics for future use in the automatic evaluation of image captions.

ACKNOWLEDGMENTS

We thank Katherine Sun, Elizabeth Salesky, and Jaimie Patterson for their valuable feedback and assistance in this work.

REFERENCES

[1] [n.d.]. Blind and Visually Impaired Assistive Technology - powered by the CloudSight.ai Image Recognition API. https://taptapseeapp.com/

[2] [n.d.]. Complex images. https://www.w3.org/WAI/tutorials/images/complex/

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*. Springer, 382–398.

[4] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7 (2019), 597–610.

[5] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv:2110.01963 [cs.CY]

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[8] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. " Nobody Speaks that Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[9] Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences. *England: Routledge* (1988).

[10] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*. 376–380.

[11] Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 452–457.

[12] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*. 549–559.

[13] Cole Gleason, Patrick Carrington, Lydia B Chilton, Benjamin Gorman, Hernisa Kacorri, Andrés Monroy-Hernández, Meredith Ringel Morris, Garreth Tigwell, and Shaomei Wu. 2020. Future research directions for accessible social media. *ACM SIGACCESS Accessibility and Computing* 127 (2020), 1–12.

[14] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

[15] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*. Springer, 417–434.

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718* (2021).

[17] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[18] Anett Hoppe, David Morris, and Ralph Ewerth. 2021. Evaluation of Automated Image Descriptions for Visually Impaired Students. In *International Conference on Artificial Intelligence in Education*. Springer, 196–201.

[19] Apple Inc. 2021. Announce notifications with Siri on airpods or beats. https://support.apple.com/en-us/HT210406

[20] Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. 2021. Communicating Visualizations without Visuals: Investigation of Visualization Alternative Text for People with Visual Impairments. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1095–1105.

[21] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating Automatic Metrics for Image Captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 199–209.

[22] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. *arXiv preprint arXiv:2106.14019* (2021).

[23] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 34–39.

[24] Jaewook Lee, Yi-Hao Peng, Jaylin Herskovitz, and Anhong Guo. 2021. Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[25] Tomer Levinboim, Ashish V Thapliyal, Piyush Sharma, and Radu Soricut. 2021. Quality Estimation for Image Captions Based on Large-scale Human Evaluations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3157–3166.

[26] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13–23.

[28] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[29] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5988–5999.

[30] Meredith Ringel Morris, Jazette Johnson, Cynthia L Bennett, and Edward Cutrell. 2018. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–11.

[31] Uran Oh, Hwayeon Joh, and YunJung Lee. 2021. Image Accessibility for Screen Reader Users: A Systematic Review and a Road Map. *Electronics* 10, 8 (2021), 953.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[33] Hermes Pique, Wieland Jeffrey, and Wu Shame. 2020. Powered by AI: Automatic alt text to help the blind 'see' Facebook. https://tech.fb.com/using-artificial-intelligence-to-help-blind-people-see-facebook/

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[36] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.

[37] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2018. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353.

[38] Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell system technical journal* 30, 1 (1951), 50–64.

[39] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. 2020. " I Hope This Is Helpful" Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[40] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. " Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[41] Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–15.

[42] Simon Thorpe, Denis Fize, and Catherine Marlot. 1996. Speed of processing in the human visual system. *nature* 381, 6582 (1996), 520–522.

[43] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 652–663. https://doi.org/10.1109/TPAMI.2016.2587640

[44] Luis Von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 79–82.

[45] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1180–1192.

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[47] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2353–2362.