# Supplementary Material

# A Sherlock Data Collection and Evaluation

The dataset was collected during the month of February of 2021. The data collected is in English and HITs were open to workers originating from US, Canada, Great Britain and Australia. We target for a worker payment rate of $15/hour for all our HITs. For data collection and qualifications, average pay for the workers came to $16-$20 with median workers being compensated $12/hour. We hash Worker IDs to preserve anonymity. A sample of data collection HIT is shown in Fig. 4 (with instructions shown in Fig. 3).

## A.1 Qualification of Workers

As a means for ensuring high quality annotations, 266 workers were manually selected through a qualification and training rounds. The workers were presented with three images and asked to provide three OBSERVATION PAIRS per image. Each of the worker responses were manually evaluated. A total 297 workers submitting 8 reasonable OBSERVATION PAIRS out of of 9 were qualified for training.

The process of creating bounding boxes and linking these boxes to the OBSERVATION PAIRS was complex enough to necessitate a training stage. For the training round, qualified workers were given a standard data collection hit (Fig. 4) at a higher pay to account for the time expected for them to learn the process. An additional training round was encouraged for a small pool of workers to ensure all workers were on the page with regards to the instructions and the mechanism of the hit. 266 workers worked on and completed the training (remaining 31 did not return for the training round). In this paper, we use the term *qualified workers* to refer to the workers who have completed both the qualification and training round.

## A.2 Data Collection

As described in §3, we collected a total of 363K OBSERVATION PAIRS which consist of a clue and inference. Further examples of annotations are shown in Fig. 7.

**Image sourcing.** For VCR images, we use the subset also annotated by VisualCOMET [44]; we limit our selection to images that contain at least 3 unique entities (persons or objects). For Visual Genome, during early annotation rounds, crowdworkers shared that particular classes of images were common and less interesting (e.g., grazing zebras, sheep in pastures). In response, we performed a semantic de-duplication step by hierarchical clustering into 80K clusters of extracted CLIP `ViT-B/32` features [51] and sample a single image from each resulting cluster. We annotate 103K images in total, and divide them into a training/validation/test set of 90K/6.6K/6.6K, aligned with the community standard splits for these corpora.

**Bounding boxes.** For each clue in an OBSERVATION PAIR, the workers were asked to draw one or more bounding boxes around image regions relevant to the clue. For example, for the clue "a lot of architectural decorations" given for the lower right image in Fig. 7, the worker chose box each of the architectural features separately in their own bounding box. While it was not strictly enforced, we encouraged the workers to keep to a maximum of 3 bounding boxes per clue, with allowance for more if necessitated by the image and the OBSERVATION PAIR, based on worker's individual discretion.

## A.3 Corpus Validation

To verify the quality of annotation, we run a validation over 17K OBSERVATION PAIRS. For each OBSERVATION PAIR, we present three independent crowdworkers with its associated image and its annotation: the clue with its corresponding region bound-boxed in the image and the inference along with its confidence rating. The workers are then asked rate the OBSERVATION PAIRS along three dimensions: (1) acceptability of the OBSERVATION PAIR (is the OBSERVATION PAIR reasonable given the image?), (2) appropriateness of bounding boxes (do the bounding boxes appropriately represent the clue?), and (3) interestingness of the OBSERVATION PAIR (how interesting is the OBSERVATION PAIR?). The annotation template of the HIT is shown in Fig. 5.

## A.4 Details on exploration of social biases

The clues and inferences we collect from crowdsource workers are abductive, and thus are uncertain. Despite this type of reasoning being an important aspect of human cognition, heuristics and assumptions may reflect false and harmful social biases. As a concrete example: early on in our collection process during a qualifying round, we asked 70 workers to annotate an image of a bedroom, where action figures were placed on the bed. Many said that the bedroom was likely to belong to a *male* child, citing the action figures as evidence. We again emphasize that our goal is to *study* heuristic reasoning, without endorsing the particular inferences themselves.

**Sample analysis.** While curating the corpus, we (the authors) have examined several thousand annotations. To supplement our qualitative experience, in addition, we conducted a close reading of a random sample of 250 inferences. This close reading was focused on references to protected characteristics of people and potentially offensive/NSFW cases.

During both our informal inspection and close reading, we observed similar patterns. Like in other vision and language corpora depicting humans, the most common reference to a protected characteristic was perceived gender, e.g., annotators often assumed depicted people were "a man" or "a woman" (and sometimes, age is also assumed, e.g., "an old man"). Aside from perception standing-in for identity, a majority of inferences are not specifically/directly about protected characteristics and are SFW (243/250 in our sample). The

small number of exceptions included: assumptions about the gender of owners of items similar to the action figure example above (1/250 cases); speculation about the race of an individual based on a sweater logo (1/250); and commenting on bathing suits with respect to gender (1/250).

Since still frames in VCR are taken from movies, some depict potentially offensive imagery, e.g., movie gore, dated tropes, etc. The images in VCR come with the following disclaimer, which we also endorse (via visualcommonsense.com): "many of the images depict nudity, violence, or miscellaneous problematic things (such as Nazis, because in many movies Nazis are the villains). We left these in though, partially for the purpose of learning (probably negative but still important) commonsense implications about the scenes. Even then, the content covered by movies is still pretty biased and problematic, which definitely manifests in our data (men are more common than women, etc.)."

**Statistical analysis.** While the random sample analysis suggests that a vast majority of annotations in our corpus do not reference protected characteristics and are SFW, for an additional check, we passed a random set of 30K samples (10K each from training/val/test) clues/inferences through the Perspective API.[1] While the API itself is imperfect and itself has biases [18, 38, 55], it nonetheless can provide some additional information on potentially harmful content in our corpus. We examined the top 50 clue/inference pairs across each split marked as most likely to be toxic. Most of these annotations were false positives, e.g., "a dirty spoon" was marked as potentially toxic likely because of the word "dirty." But, this analysis did highlight a very small amount of lewd/NSFW/offensive content. Out of the 30K cases filtered through the perspective API, we discovered 6 cases of weight stigmatization, 2 (arguably) lewd observation, 1 dark comment about a cigarette leading to an early death for a person, 1 (arguable) case of insensitivity to mental illness, 6 cases of sexualized content, and 1 (arguable) case where someone was highlighted for wearing non-traditionally-gendered clothing.

# B Additional Modeling Details

After some light hyperparameter tuning on the validation set, the best learning rate for fine-tuning our CLIP models was found to be .00001 with AdamW [35, 27]. We use a linear learning rate warmup over 500 steps for `RN50x16` and `ViT-B/16`, and 1000 for `RN50x64`. Our biggest model, `RN50x64`, takes about 24 hours to converge when trained on 8 Nvidia RTX6000 cards. For data augmentation during training, we use `pytorch`'s `RandomCrop`, `RandomHorizontalFlip`, `RandomGrayscale`, and `ColorJitter`. For our widescreen CLIP variants, data augmentations are executed on each half of the image independently. We compute visual/textual embeddings via a forward pass of the respective branches of CLIP — for our widescreen model, we simply average the resultant embeddings

[1] `https://www.perspectiveapi.com/`; November 2021 version.

| | *Retrieval* | | *Localization* |
|---|---|---|---|
| | im → txt (↓) | $P@1_{im \to txt}$ (↑) | GT-Box/Auto-Box (↑) |
| `RN50x64`-inference | 12.8 | 43.4 | 92.5/41.4 |
| `RN50x64`-clue | 6.2 | 54.3 | 94.7/53.3 |
| `RN50x64`-multitask | **5.4** | **57.5** | **95.3/54.3** |

Table 1: Retrieval and localization results when clues are used at evaluation time instead of inferences. This task is more akin to referring expression retrieval/localization rather than abductive commonsense reasoning. While clue retrieval/localization setups are easier overall (i.e., referring expressions are easier both models to reason about) the model trained for abductive reasoning, `RN50x64`-inference, performs worse than the model trained on referring expressions `RN50x64`-clue.

for each side of the image. To compute similarity score, we use cosine similarity, and then scale the resulting similarities using a logit scaling factor, following [51]. Training is checkpointed every 300 gradient steps, and the checkpoint with best validation $P@1$ retrieval performance is selected.

**Ablation details.** For all ablations, we use the `ViT-B/16` version of CLIP for training speed: this version is more than twice as fast as our smallest ResNet, and enabled us to try more ablation configurations.

**A cleaner training corpus.** Evaluations are reported over version 1.1 of the Sherlock validation/test sets. However, our models are trained on version 1.0, which contains 3% more data; early experiments indicate that the removed data doesn't significantly impact model performance. This data was removed because we discovered a small number of annotators were misusing the original collection interface, and thus, we removed their annotations. We encourage follow-up work to use version 1.1, but include version 1.0 for the sake of replicability.

**T5 model details.** We train T5-Large to map from clues to inferences using the Huggingface transformers library [68]; we parallelize using the Huggingface `accelerate` package. We use Adafactor [58] with learning rate .001 and batch size 32, train for 5 epochs, and select the checkpoint with the best validation loss.

## B.1 Results on Clues instead of Inferences

Whereas inferences capture abductive inferences, clues are more akin to referring expressions. While inferences are our main focus at evaluation time, **Sherlock** also contains an equal number of clues, which act as literal descriptions of image regions: **Sherlock** thus provides a new dataset of 363K localized referring expressions grounded in the image regions of VisualGenome and VCR. As a

pointer towards future work, we additionally report results for the retrieval and localization setups, but instead of using a version testing on inference texts, we test on clues. We do not report over our human-judged comparison sets, because or raters only observed inferences in that case. Table 1 includes prediction results of two models in this setting: both are `RN50x64` models trained with widescreen processing and with clues highlighted in pixel space, but one is trained on inferences, and one is trained on clues.

# C Batch Size Ablation

We hypothesize the nature of the hard negatives the models encounter during training is related to their performance. Because UNITER and LXMERT are bidirectional, they are quadratically more memory intensive vs. CLIP: as a result, for those models, we were only able to train with 18 negative examples per positive (c.f. CLIP `ViT-B/16`, which uses 511 negatives). To check that batch size/number of negatives wasn't the only reason CLIP outperformed UNITER, we conducted an experiment varying `ViT-B/16`'s batch size from 4 to 512; the results are given in Fig. 1. Batch size doesn't explain all performance differences: with a batch size of only 4, our weakest CLIP-based model still localizes better than UNITER, and, at batch size 8, it surpasses UNITER's retrieval performance.

# D Clues and inferences vs. literal captions

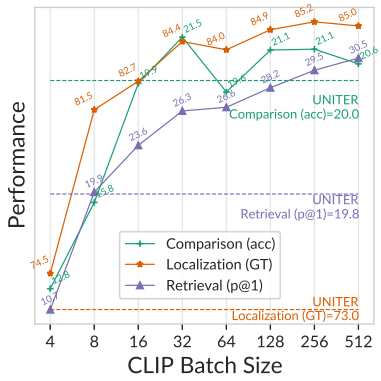


Figure 1: The effect of batch size on performance of `ViT/B-16`. UNITER batch size is 256. Performance on all tasks increases with increasing batch size, but appears to saturate, particularly for comparison.

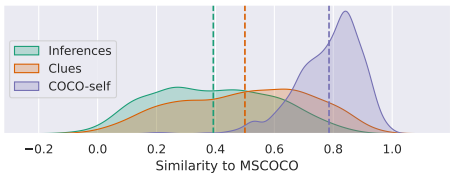


Figure 2: The SentenceBERT [53] cosine similarity between clues/inferences and MSCOCO captions; MSCOCO caption self-similarity included for reference. On average, clues are closer to MSCOCO captions than inferences.

We ran additional analyses to explore the textual similarity between **Sherlock**'s clues and inferences vs. literal image descriptions. For 2K images, we computed text overlap using S-BERT cosine similarity [53] between MS COCO captions and **Sherlock** clues/inferences. The result is in Fig. 2. As a baseline we include COCO self-similarity with held-out captions. Clues are more similar to COCO captions than inferences, presumably because they make reference to the same types of literal objects/actions that are described in literal captions.

# E Comparison Human Evaluation Set Details

We aim to sample a diverse and plausible set of candidate inferences for images to form our comparison set. Our process is a heuristic effort designed to elicit "interesting" annotations from human raters. Even if the process isn't perfect for generating interesting candidates, because we solicit human ratings we show inferences to annotators and ask them to rate their plausibility, the resulting set will still be a valid representation of human judgment. We start by assuming all inferences could be sampled for a given image+region, and proceed to filter according to several heuristics.

First, we use a performant `RN50x16` checkpoint as a means of judging plausibility of inferences. This checkpoint achieves 18.5/20.6/31.5 im2txt/txt2im/P@1 respectively on retrieval on v1.0 of the **Sherlock** corpus; this is comparable to the `RN50x16` checkpoint we report performance on in our main results section. We use this checkpoint to score all validation/test (image+region, inference) possibilities.

**Global filters.** We assume that if the model is already retrieving its ground truth inference which high accuracy, the instance is probably not as interesting: for each image, we disqualify all inferences that receive a lower plausibility estimate from our `RN50x16` checkpoint vs. the ground truth inference (this also discards the ground-truth inference). This step ensures that the negative inferences we sample are more plausible than the ground truth inference according to the model. Next, we reduce repetitiveness of our inference texts using two methods. First, we perform the same semantic de-duplication via hierarchical clustering as described in § 3: clustering is computed on SentenceBERT [53] representations of inferences (`all-MiniLM-L6-v2`). We compute roughly 18K clusters (corresponding to 80% of the dataset size) and sample a single inference from each cluster: this results in 20% of the corpus being removed from consideration, but maintains diversity, because each of the 18K clusters is represented. Second, we perform a hard-deduplication by only allowing three verbatim copies of each inference to be sampled.

**Local filters.** After these global filters, we begin the iterative sampling process for each image+region. If, after all filtering, a given image+region has fewer than 20 candidates to select from, we do not consider it further. Then, in a greedy fashion, we build-up the candidate set by selecting the remaining

inference with i) the highest model plausibility ii) that is maximally dissimilar to the already sampled inferences for this image according to the Sentence-BERT representations. Both of these objectives are cosine similarities in vector spaces (one between image and text, and one between text and text). We assign weights so that the image-text similarity (corresponding to `RN50x16` plausibility) is 5x more important than the text-text dissimilarity (corresponding to SentenceBERT diversity). After iteratively constructing a diverse and plausible set of 10 inferences for a given image under this process, we globally disqualify the sampled inferences such that no inference is sampled more than once for each image (unless it is a verbatim duplicate, in which case, it may be sampled up to 3 times).

Finally, for all of the images we are able to sample a set of 10 inferences for, we sort by how promising they are collectively according to a weighted sum of: the (globally ranked) average length of the sampled inferences, the (globally ranked) diversity of the set of 10 (measured by mean all-pairs SentenceBERT cosine sim: lower=more diverse), and 5x the (globally ranked) average plausibility according to `RN50x16`. We collect 2 human judgments for each of the 10 inferences for the top 500 images from the val/test sets (1K total) according to this heuristic ranking. The total is 20K human judgments, which formed v1 of the **Sherlock** comparison corpus. v1.1 has 19K judgments.

**Crowdowrking details.** For the comparison task, we designed an additional HIT to collect human feedback on the retrieved inferences. In the HIT, workers were presented with the images with the appropriate clue region highlighted. Then they were provided with the inferences and were asked to rate them on a likert scale of 1-3, with 1 as "irrelevant" or "verifiably incorrect", 2 as "statement is probably true but there is a better highlighted region to support it", and 3 as "statement is probably true and the highlighted region supports it". A sample of evaluation HIT is shown in Fig. 6. Human agreement on this setup is reported as accuracy §5.1.

# F Datasheet for Sherlock

In this section, we present a Datasheet [14, 4] for **Sherlock**.

1. Motivation For Datasheet Creation

   - **Why was the dataset created? Sherlock** was created to support the study of visual abductive reasoning. Broadly speaking, in comparison to corpora which focus on concrete, objective facets depicted within visual scenes (e.g., the presence/absence of objects), we collected **Sherlock** with the goal of better understanding the types of abductive inferences that people make about images. All abductive inferences carry uncertainty. We aim to study the inferences we collect, but do not endorse their objectivity, and do not advocate for use cases that risk perpetuating them.

- **Has the dataset been used already?** The annotations we collect are novel, but the images are sourced from two widely-used, existing datasets: Visual Genome [29] and VCR [75].
- **What (other) tasks could the dataset be used for?** Aside from our retrieval/localization setups, **Sherlock** could be useful as a pretraining corpus for models that aim to capture information about what people might assume about an image, rather than what is literally depicted in that image. One potentially promising case: if a malicious actor were posting emotionally manipulative content online, it might be helpful to study the types of assumptions people might make about their posts, rather than the literal contents of the post itself.
- **Who funded dataset creation?** This work was funded by DARPA MCS program through NIWC Pacific (N66001-19-2-4031), the DARPA SemaFor program, and the Allen Institute for AI.

2. Data composition

- **What are the instances?** We refer to the instances as clues/inferences, which are authored by crowdworkers. As detailed in the main text of the paper, a clue is a bounding box coupled with a free-text description of the literal contents of that bounding box. An inference is an abductive conclusion that the crowdworker thinks could be true about the clue.
- **How many instances are there?** There are 363K commonsense inferences grounded in 81K Visual Genome images and 22K VCR images.
- **What data does each instance consist of?** Each instance contains 3 things: a clue, a short English literal description of a portion of the image, an inference, a short English description of an inference associated with the clue that aims to be not immediately obvious from the image content, and a bounding box specified with the region of interest.
- **Is there a label or target associated with each instance?** We discuss in the paper several tasks, which involve predicting inferences, bounding boxes, etc.
- **Is any information missing from individual instances?** Not systematically — in rare circumstances, we had to discard some instances because of malformed crowdworking inputs.
- **Are relationships between individual instances made explicit?** Yes — the annotations for a given image are all made by the same annotator and are aggregated based on that.
- **Does the dataset contain all possible instances or is it a sample?** This is a natural language sample of abductive inferences; it would probably be impossible to enumerate all of them.
- **Are there recommended data splits?** Yes, they are provided.
- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** Yes: some annotations are repeated by crowdworkers. When we collected the corpus of Likert judgments for evaluation, we performed both soft and hard deduplication steps, ensuring that the text people were evaluating wasn't overly repeti-

tive.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** It links to the images provided by Visual Genome and VCR. If images were removed from those corpora, our annotations wouldn't be grounded.

3. Collection Process

- **What mechanisms or procedures were used to collect the data?** We collected data using Amazon Mechanical Turk.
- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred or derived from other data?** Paid crowdworkers provided the annotations.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** We downsample common image types via a semantic deduplication step. Specifically, some of our crowdworkers were rightfully pointing out that it's difficult to say interesting things about endless pictures of zebra; these types of images are common in visual genome. So, we performed hierarchical clustering on the images from that corpus, and then sampled 1 image from each of 80K clusters. The result is a downsampling of images with similar feature representations. We stopped receiving comments about zebras after this deduplication step.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** Crowdworkers constructed the corpus via a mechanical turk HIT we designed. We our target was to pay $15/hour. A post-hoc analysis revealed that crowdworkers were paid a median $12/hr and a mean of $16-20/hour, depending on the round.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** The main data was collected in February 2021.

4. Data Preprocessing

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Yes, significant preprocessing was conducted. The details are in
- **Was the "raw" data saved in addition to the preprocessed, cleaned, labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the 'raw' data.** The concept of "raw" data is difficult to specify in our case. We detail the data

we release in the main body of the paper.

- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** We plan to release some software related to modeling, and also have provided some appendices that detail the crowdworking labelling efforts.
- **Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?** We think so. It's difficult to fully specify the abductive reasoning process of humans. But we think our work goes a step beyond existing corpora.

5. Dataset Distribution

   - **How will the dataset be distributed?**

     The dataset is available at `http://visualabduction.com/`.
   - **When will the dataset be released/first distributed? What license (if any) is it distributed under?**

     The dataset is released under CC-BY 4.0 and the code is released under Apache 2.0.
   - **Are there any copyrights on the data?**

     The copyright for the new annotations is held by AI2 with all rights reserved.
   - **Are there any fees or access restrictions?**

     No — our annotations are freely available.

6. Dataset Maintenance

   - **Who is supporting/hosting/maintaining the dataset?**
     The dataset is hosted and maintained by AI2.
   - **Will the dataset be updated? If so, how often and by whom?**
     We do not currently have plans to update the dataset regularly.
   - **Is there a repository to link to any/all papers/systems that use this dataset?**
     No, but if future work finds this work helpful, we hope they will consider citing this work.
   - **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?**
     People are free to remix, use, extend, build, critique, and filter the corpus: we would be excited to hear more about use cases either via our github repo, or via personal correspondence.

7. Legal and Ethical Considerations

   - **Were any ethical review processes conducted (e.g., by an institutional review board)?**
     Crowdworking studies involving no personal disclosures of standard computer vision corpora are not required by our IRB to be reviewed

by them. While we are not lawyers, the opinion is based on United States federal regulation 45 CFR 46, under which this study qualifies and as exempt and does not require IRB review.

(a) We do not collect personal information. Information gathered is strictly limited to general surveys probing at general world knowledge.

(b) We take precaution to anonymize Mechanical WorkerIDs in a manner that the identity of the human subjects cannot be readily ascertained (directly or indirectly).

(c) We do not record or include any interpersonal communication or contact between investigation and subject.

Specifically:

- We do not have access to the underlying personal records and will record information in such a manner that the identity of the human subject cannot readily be ascertained.
- Information generated by participants is non-identifying without turning over the personal records attached to these worker IDs.
- We do not record or include any interpersonal communication or contact between investigation and subject.

- **Does the dataset contain data that might be considered confidential?**

  Potentially, yes. Most of the content in the corpus that would be considered potentially private/confidential would likely be depicted in the images of Visual Genome (VCR are stills from movies where actors on-screen are presumably aware of their public actions). While we distribute no new images, if an image is removed from Visual Genome (or VCR), it will be removed from our corpus as well.

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

  As detailed in the main body of the paper, we have searched for toxic content using a mix of close reading of instances and the Perspective API from Google. In doing this, we have identified a small fraction of instances that could be construed as offensive. For example, in a sample of 30K instances, we discovered 6 cases that arguably offensive (stigmatizes depicted people's weight based on visual cues). Additionally, some of the images from VCR, gathered from popular movies, can depict potentially offensive/disturbing content. The screens can be "R Rated," e.g., some images depict movie violence with zombies, some of the movies have Nazis as villains, and thus, some of the screenshots depict Nazi symbols. We reproduce VCR's content warning about such imagery in § A.2.

- **Does the dataset relate to people?**

  Yes: the corpus depicts people, and the annotations are frequently abductive inferences that relate to people. As detailed in the main body of the paper, 36% of inferences (or more) are grounded on people; and, many inferences that are not directly grounded on people may relate to them. Moreover, given that we aim to study abduction, which is an intrinsically subjective process, the annotations themselves are, at least in part, reflections of the annotators themselves.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?**

  We don't explicitly disallow identification by gender or age, e.g., in the clues/inferences, people often will use gendered pronouns or aged language in reference to people who are depicted (e.g., "the old man"). Furthermore, while we undertook the sample/statistical toxicity analysis detailed in the main body of the paper, we have not manually verified that all 363K clue/inference pairings are free of any reference to a subpopulation. For example, we observed one case wherein an author speculated about the country-of-origin of an individual being Morroco, clued by the observation that they were wearing a fez. Like the other observations in our corpus, it's not necessarily the case that this is an objectively true inference, even if the fez is a hat that is worn in Morroco.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

  The data collection process specifically instructs workers to avoid identifying any individual in particular (e.g., actors in movie scenes). Instead, they are specifically instructed to use general identifiers to describe people (e.g. "student", "old man", "engineer"). In our experience with working with the corpus, we haven't encountered any instances where our annotators specifically identified anyone, e.g., by name. The images contained in VCR and Visual Genome that we source from do contain uncensored images of faces. But, if images are removed from those corpora, they will be removed from Sherlock as well, as we do not plan to re-host the images ourselves.

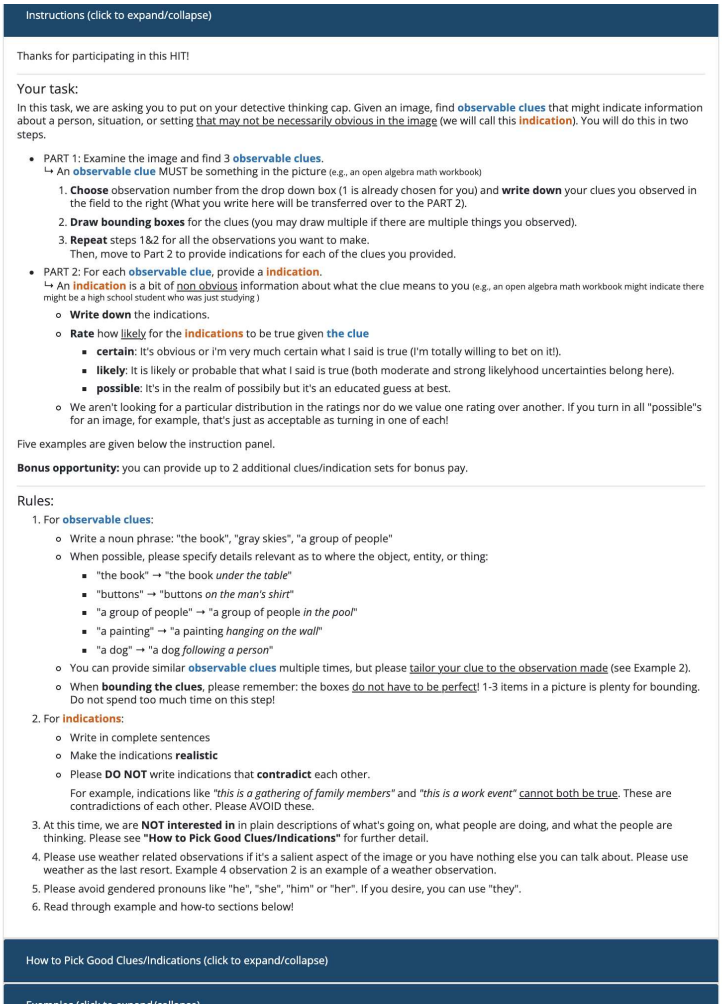


Figure 3: Instructions for **Sherlock** data collection HIT.


Figure 4: Template setup for **Sherlock** data collection HIT. Instructions are shown in Figure 3

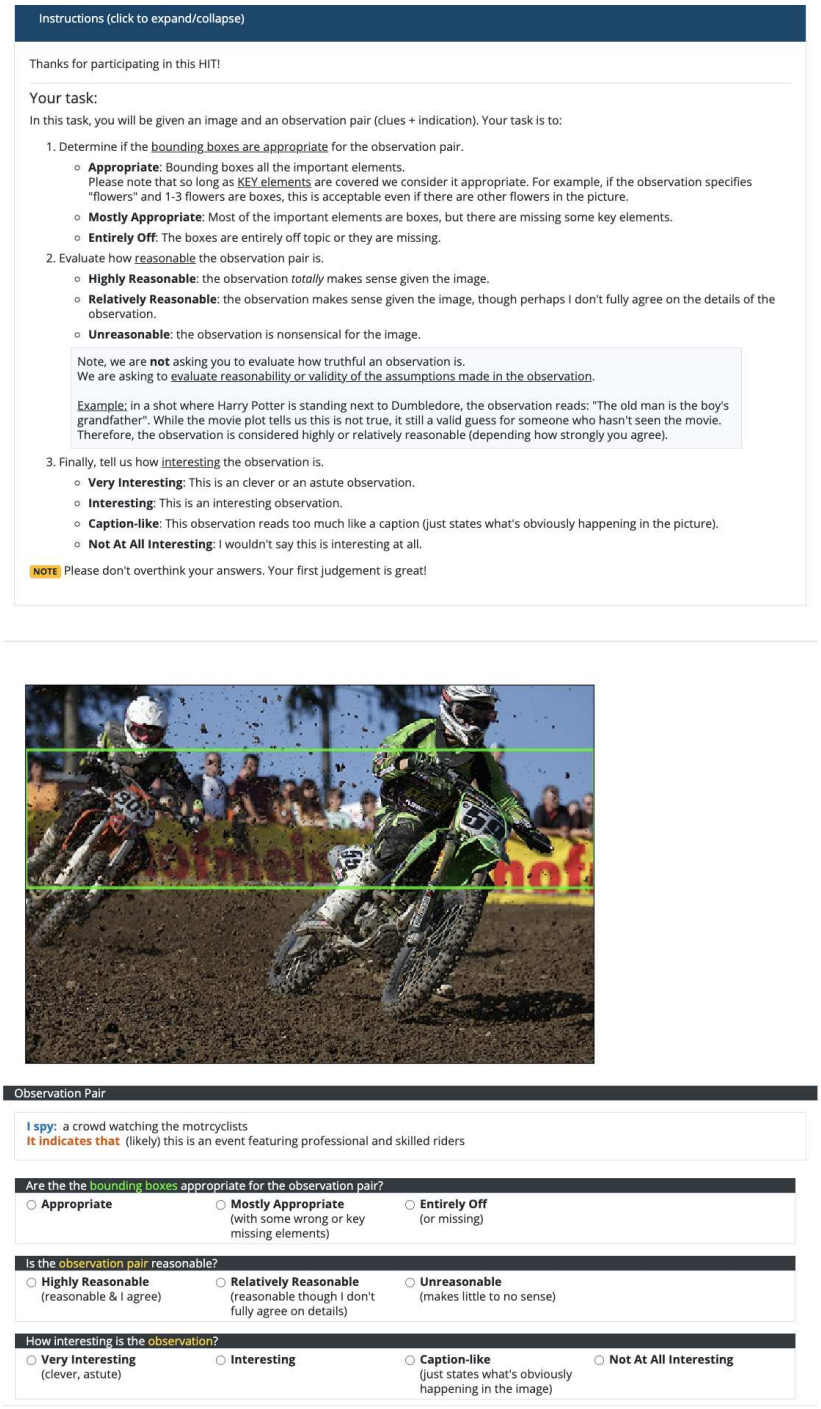


Figure 5: Instructions and template setup for **Sherlock** data validation HIT.

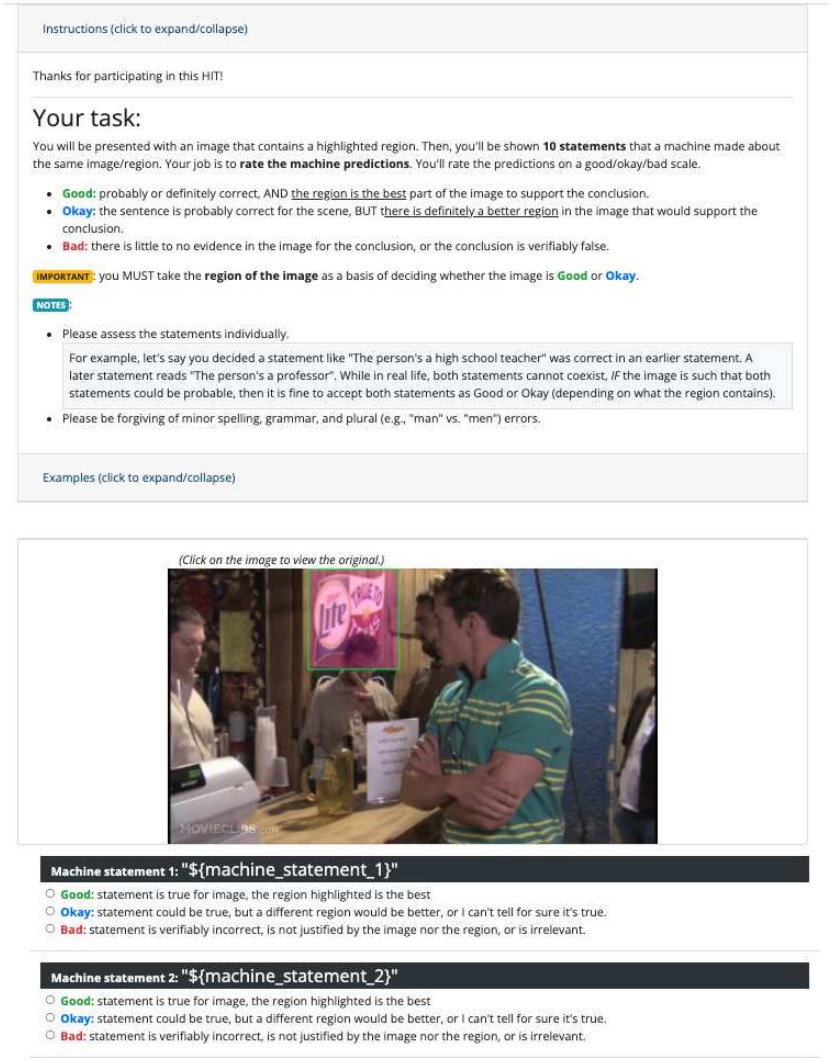


Figure 6: Instructions and template setup for **Sherlock** model evaluation HIT.

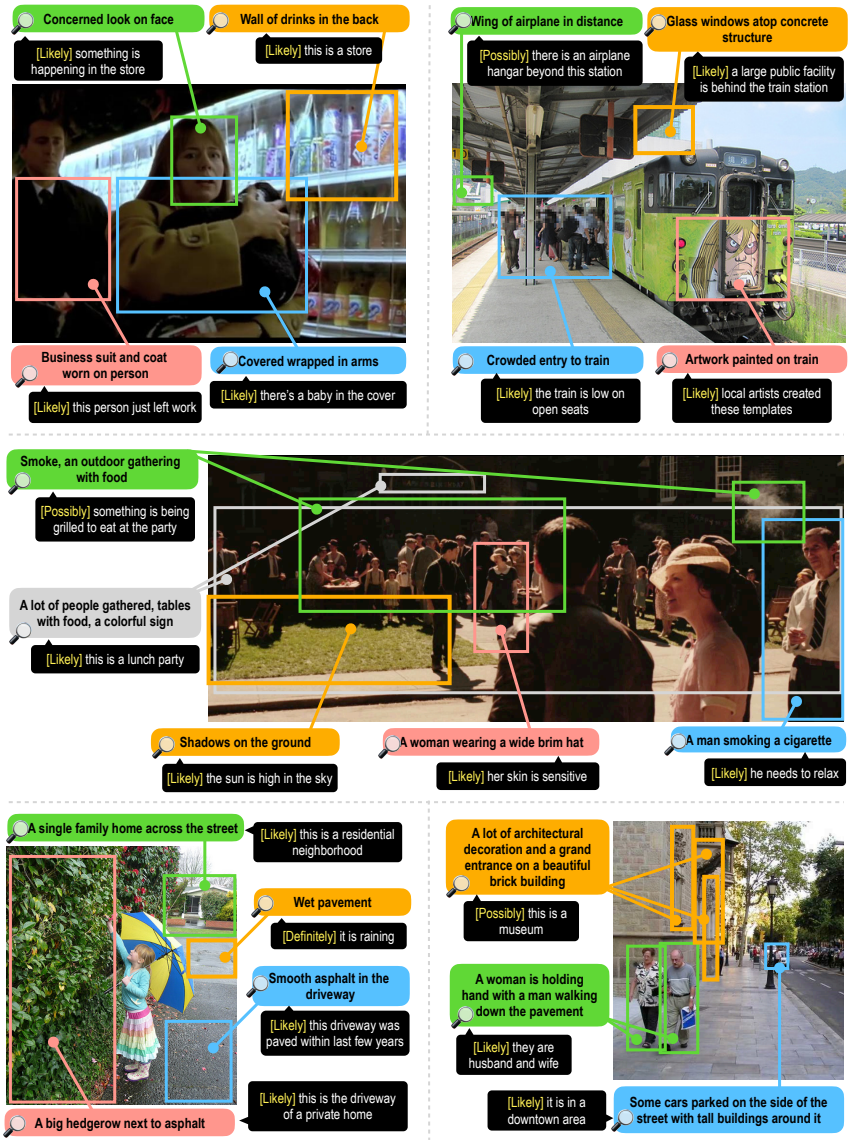


Figure 7: Examples of clues and inference pair annotations in **Sherlock** over images from Visual Genome and VCR. For each OBSERVATION PAIR, an inference (speech bubble) is grounded in a concrete clue (color bubble) present in an image. CONFIDENCE SCORE (in the order of decreasing confidence: "Definitely" > "Likely" > "Possibly") for each inference is shown in yellow.