The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning

Jack Hessel^{*1}, Jena D. Hwang^{*1}, Jae Sung Park², Rowan Zellers², Chandra Bhagavatula¹, Anna Rohrbach³, Kate Saenko⁴, and Yeiin Choi^{1,2}

¹ Allen Institute for AI {jackh, jenah, chandrab}@allenai.org

 $^2\,$ Paul G. Allen School of Computer Science & Engineering, University of Washington

{jspark96,rowanz,yejin}@cs.washington.edu

³ University of California, Berkeley anna.rohrbach@berkeley.edu

 $^4\,$ Boston University and MIT-IBM Watson AI saenko@bu.edu

Abstract. Humans have remarkable capacity to reason abductively and hypothesize about what lies beyond the literal content of an image. By identifying concrete visual **clues** scattered throughout a scene, we almost can't help but draw probable **inferences** beyond the literal scene based on our everyday experience and knowledge about the world. For example, if we see a "20 mph" sign alongside a road, we might assume the street sits in a residential area (rather than on a highway), even if no houses are pictured. Can machines perform similar visual reasoning?

We present Sherlock, an annotated corpus of 103K images for testing machine capacity for abductive reasoning beyond literal image contents. We adopt a free-viewing paradigm: participants first observe and identify salient clues within images (e.g., objects, actions) and then provide a plausible inference about the scene, given the clue. In total, we collect 363K (clue, inference) pairs, which form a first-of-its-kind abductive visual reasoning dataset. Using our corpus, we test three complementary axes of abductive reasoning. We evaluate the capacity of models to: i) retrieve relevant inferences from a large candidate corpus; ii) *localize* evidence for inferences via bounding boxes, and iii) compare plausible inferences to match human judgments on a newlycollected diagnostic corpus of 19K Likert-scale judgments. While we find that fine-tuning CLIP-RN50x64 with a multitask objective outperforms strong baselines, significant headroom exists between model performance and human agreement. Data, models, and leaderboard available at http://visualabduction.com/.

> You know my method. It is founded upon the observation of trifles.

"The Boscombe Valley Mystery", by A. C. Doyle



Fig. 1: We introduce **Sherlock**: a corpus of 363K commonsense inferences grounded in 103K images. Annotators highlight localized clues (color bubbles) and draw plausible abductive inferences about them (speech bubbles). Our models are able to predict localized inferences (top predictions are shown), but we quantify a large gap between machine performance and human agreement.

1 Introduction

The process of making the most plausible inference in the face of incomplete information is called *abductive reasoning*, [47] personified by the iconic visual inferences of the fictional detective Sherlock Holmes.⁵ Upon viewing a scene, humans can quickly synthesize cues to arrive at abductive hypotheses that go beyond the what's captured in the frame. Concrete cues are diverse: people take into account the emotion and mood of the agents, speculate about the rationale for the presence/absence of objects, and zero-in on small, contextual details; all the while accounting for prior experiences and (potential mis)conceptions.⁶ Fig. 1 illustrates: snow may imply dangerous road conditions, an Ohio licence plate may suggest the location of the accident, and a blue sign may indicate this road is an interstate. Though not all details are equally important, certain *salient* details shape our abductive inferences about the scene as a whole [56]. This type of visual information is often left unstated.

We introduce **Sherlock**, a new dataset of 363K commonsense inferences grounded in 103K images. **Sherlock** makes explicit typically-unstated cognitive processes: each image is annotated with at least 3 inferences which pair depicted details (called clues) with commonsense conclusions that aim to go *beyond* what is literally pictured (called inferences). **Sherlock** is more diverse than many existing visual commonsense corpora like Visual Commonsense Reasoning [75]

⁵ While Holmes rarely makes mistakes, he frequently misidentifies his mostly abductive process of reasoning as "deductive." [39,8]

⁶ The *correctness* of abductive reasoning is certainly not guaranteed. Our goal is to study perception and reasoning *without* endorsing specific inferences (see §3.1).

Dataset	# Images	Format	bboxes?	free- viewing?	human- centric?
VCR [75]	110K	QA	~		~
VisualCOMET [44]	59K	If/Then KB	\checkmark		\checkmark
Visual7W [79]	47K	QA	\checkmark	partial	
Visual Madlibs [72]	11K	FiTB	\checkmark	partial	√
Abstract Scenes [65]	4.3K	KB			
Why In Images [49]	792	KB			\checkmark
BD2BB [48]	3.2K	If/Then		√	√
FVQA [66]	2.2K	QA+KB			
OK-VQA [36]	14K	QA		√	
KB-VQA [67]	700	QA	√		
Sherlock	103K	clue/inference	• √	✓	

Table 1: Comparison between **Sherlock** and prior annotated corpora addressing visual abductive reasoning from static images. **Sherlock** showcases a unique data collection paradigm, leading to a rich variety of non-human centric (i.e., not solely grounded in human references) visual abductive inferences.

and VisualCOMET [44],⁷ due to its free-viewing data collection paradigm: we purposefully do not pre-specify the types of clues/inferences allowed, leaving it to humans to identify the most salient and informative elements and their implications. Other forms of free-viewing like image captions may not be enough: a typical caption for Fig. 1 may mention the accident and perhaps the snow, but smaller yet important details needed to comprehend the larger scene (like the blue freeway sign or the Ohio plates) may not be mentioned explicitly [5]. Dense captioning corpora [22] attempts to overcome this problem by highlighting *all* details, but it does so without accounting for which details are salient (and why).

Using our corpus, we propose three complementary tasks that evaluate different aspects of machine capacity for visual abductive reasoning:

- 1. *Retrieval of Abductive Inferences:* given an image+region, the algorithm scores a large set of candidate inferences and is rewarded for assigning a high score to the gold annotation.
- 2. Localization of Evidence: the algorithm selects a bounding box within the image that provides the best evidence for a given inference.
- 3. Comparison of Plausibility: the algorithm scores a small set of plausible inferences for a given image+region, and is rewarded for aligning its scores with human judgments over those sets.

In our setup, a single model undertakes all of these tasks: we ask algorithms to score the plausibility of an inference given an image and a bounding box contained within it.⁸ We can directly compare models in their capacity to perform abductive reasoning, without relying on indirect generation evaluation metrics.

Model predicted inferences are given in Fig. 1. The model is a fine-tuned CLIP [51] augmented to allow bounding boxes as input, enabling users to specify particular regions for the model to make abductive inferences about. Our best model, a multitask version of CLIP RN50x64, outperforms strong baselines like UNITER [9] and LXMERT [61] primarily because it pays specific attention to the

⁷ For instance, 94% of visual references in [75] are about depicted actors, and [44] even requires KB entries to explicitly regard people; see Fig. 2.

⁸ We reserve generative evaluations (e.g., BLEU/CIDEr) for future work: shortcuts (e.g., outputting the technically correct "this is a photo" for all inputs) make generation evaluation difficult in the abductive setting (see §6). Nonetheless, generative *models* can be evaluated in our setup; we experiment with one in §5.1.



Fig. 2: Side-by-side comparison of VCR [75], VisualCOMET [44], and **Sherlock** on a representative instance. **Sherlock** showcases a wider range of (non-human centric) situational contexts.

correct input bounding box. We additionally show that 1) for all tasks, reasoning about the full context of the image (rather than just the region corresponding to the clue) results in the best performance; 2) a text-only model cannot solve the comparison task even when given oracle region descriptions; and 3) a multi-task model fit on both clues/inferences at training time performs best even when only inferences are available at test time.

We foresee **Sherlock** as a difficult diagnostic benchmark for vision-andlanguage models. On our comparison task, in terms of pairwise accuracy, our best model lags significantly below human agreement (headroom also exists for retrieval and localization). We release code, data, and models at http: //visualabduction.com/.

2 Related Work

Abductive reasoning. Abduction, a form of everyday reasoning first framed byPeirce, [46,47]; involves the creating of explanatory hypothesesbased on limited evidence. Humans use abduction to reconcile seemingly disconnected observations to arrive at meaningful conclusions [56] but readily retract in presence of new evidence [1]. In linguistics, abduction for communicated meaning (in an impoverished conversational context) is systematized through conversational maxims [15]. In images, [5] show that different object types have different likelihoods of being mentioned in image captions (e.g., "fireworks" is always mentioned if depicted, but "fabric" is not), but that object type alone does not dictate salience for abductive inferences, e.g., a TV in a living room may not be as conceptually salient as a TV in a bar, which may signal a particular type of bar. Abductive reasoning has recently received attention in language processing tasks [6,50,11,45], proof writing [60], and discourse processing [17,42], etc.

Beyond visual recognition. Several tasks that go beyond image description/recognition have been proposed, including visual and analogical reasoning [43,77,21,3], scene semantics [23], commonsense interactions [65,49], temporal/causal reasoning [26,71], and perceived importance [5]. Others have explored commonsense reasoning tasks posed over videos, which usually have more input available than a single frame [63,20,31,74,13,32,78,12,34,19] (inter alia).

Visual abductive reasoning. Sherlock builds upon prior grounded visual abductive reasoning efforts (Table 1). Corpora like Visual Commonsense Reasoning (VCR) [75], VisualCOMET [44], and Visual7W [79] are most similar to Sherlock in providing benchmarks for rationale-based inferences (i.e., the why and how). But, Sherlock differs in format and content (Fig. 2). Instead of annotated QA pairs like in [79,75] where one option is definitively correct, free-text clue/inference pairs allow for broader types of image descriptions, lending itself to softer and richer notions of reasoning (see §4)—inferences are not definitively correct vs. incorrect, rather, they span a range of plausibility. Deviating from the constrained, human-centric annotation of [44], Sherlock clue/inference pairs support a broader range of topics via our open-ended annotation paradigm (see §3). Sherlock's inferences can be grounded on any number of visual objects in an image, from figures central to the image (e.g., persons, animals, objects) to background cues (e.g., time, location, circumstances).

3 Sherlock Corpus

The **Sherlock** corpus contains a total of 363K abductive commonsense inferences grounded in 81K Visual Genome [29] images (photographs from Flickr) and 22K Visual Commonsense Reasoning (VCR) [75] images (still-frames from movies). Images have an average of 3.5 OBSERVATION PAIRS, each consisting of: - **clue**: an observable entity or object in the image, along with bounding box(es)

- specifying it (e.g., "people wearing nametags").
- **inference**: an abductive inference associated with the clue; not immediately obvious from the image content (e.g., "the people don't know each other").

Both clues and inferences are represented via free text in English; both have an average length of seven tokens; per clue, there are a mean/median of 1.17/1.0 bounding boxes per clue. We divide the 103K annotated images into a training/validation/test set of 90K/6.6K/6.6K. Further details are available in §A.

Annotation process. We crowdsource our dataset via Amazon Mechanical Turk (MTurk). For each data collection HIT, a manually qualified worker is given an image and prompted for 3 to 5 OBSERVATION PAIRS. For each OBSER-VATION PAIR, the worker is asked to write a clue, highlight the regions in the image corresponding to the clue, and write an inference triggered by the clue. To discourage purely deductive reasoning, the workers are actively encouraged to think beyond the literally depicted scene, while working within real-world expectations. Crowdworkers also self-report Likert ratings of confidence in the correctness of their abductive inferences along a scale of "definitely" = 3/3, "likely" = 2/3, and "possibly" = 1/3. The resulting inferences span this range (31%, 51%, 18%, respectively). To validate corpus quality, we run a validation round for 17K OBSERVATION PAIRS in which crowdworkers provide ratings for acceptability (is the annotation reasonable?), bboxes (are the boxes reasonably placed for the clue?), and *interestingness* (how interesting is the annotation?). We find that 97.5% of the OBSERVATION PAIRS are acceptable with 98.3% accurate box placement; and 71.9% of inferences are found to be interesting.



Fig. 3: Overview of the topics represented in the clues and inferences in **Sherlock**. This analysis shows that **Sherlock** covers a variety of topics commonly accessible in the natural world. Color of the connections reflect the clue topic.

3.1 Dataset Exploration

Sherlock's abductive inferences cover a wide variety of real world experiences from observations about unseen yet probable details of the image (e.g., "smoke at an outdoor gathering" \rightarrow "something is being grilled") to elaborations on the expected social context (e.g., "people wearing nametags" \rightarrow "[they] don't know each other"). Some inferences are highly likely to be true (e.g., "wet pavement" \rightarrow "it has rained recently"); others are less definitively verifiable, but nonetheless plausible (e.g., "large trash containers" \rightarrow "there is a business nearby"). Even the inferences crowdworkers specify as 3/3 confident are almost always abductive, e.g., wet pavement strongly but not always indicate rain. Through a rich array of natural observations, **Sherlock** provides a tangible view into the abductive inferences people use on an everyday basis (more examples in Fig. 7).

Assessing topic diversity. To gauge the diversity of objects and situations represented in **Sherlock**, we run an LDA topic model [7] over the OBSERVA-TION PAIRS. The topics span a range of common everyday objects, entities, and situations (Fig. 3). Inference topics associated with the clues include withincategory associations (e.g., "baked potatoes on a ceramic plate" \rightarrow "this [is] a side dish") and cross-category associations (e.g., "a nametag" (attire) \rightarrow "she works here" (characterization)). Many topics are not human centric; compared to VCR/VisualCOMET in which 94%/100% of grounded references are to people. A manual analysis of 150 clues reveals that only 36% of **Sherlock** OBSERVATION PAIRS are grounded on people.

Intended use cases. We manually examine of 250 randomly sampled OBSER-VATION PAIRS to better understand how annotators referenced protected characteristics (e.g., gender, color, nationality). A majority of inferences (243/250) are not directly about protected characteristics, though, a perceived gender is often made explicit via pronoun usage, e.g., "she is running." As an additional check, we pass 30K samples of our corpus through the Perspective API.⁹ A manual examination of 150 cases marked as "most toxic" reveals mostly false positives (89%), though 11% of this sample do contain lewd content (mostly prompted by

⁹ https://www.perspectiveapi.com/; November 2021 version. The API (which itself is imperfect and has biases [18,38,55]) assigns toxicity value 0-1 for a given input text. Toxicity is defined as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion."



(a) Retrieval of abductive (b) Localization of evidence (c) Comparison of plausibilinferences ity

Fig. 4: We pose three tasks over **Sherlock**: In *retrieval*, models are tasked with finding the ground-truth inference across a wide range of inferences, some much more plausible/relevant than others. In *localization*, models must align regions within the same image to several inferences written about that image. For *comparison*, we collect **19K Likert ratings** from human raters across **plausible candidates**, and models are evaluated in their capacity to reconstruct human judgments across the candidates. Despite intrinsic subjectivity, headroom exists between human agreement and model performance, e.g., on the *comparison* task.

visual content in the R-rated VCR movies) or stigmas related to, e.g., gender and weight. See §A.4 for a more complete discussion.

While our analysis suggests that the relative magnitude of potentially offensive content is low in **Sherlock**, we still advocate against deployed use-cases that run the risk of perpetuating potential biases: our aim is to study abductive reasoning without endorsing the correctness or appropriateness of particular inferences. We foresee **Sherlock** as 1) a diagnostic corpus for measuring machine capacity for visual abductive reasoning; 2) a large-scale resource to study the types of inferences people may make about images; and 3) a potentially helpful resource for building tools that require understanding *abductions* specifically, e.g., for detecting purposefully manipulative content posted online, it could be useful to specifically study what people *might assume* about an image (rather than what is objectively correct; more details in Datasheet (\S F) [14]).

4 From Images to Abductive Inferences

We operationalize our corpus with three tasks, which we call retrieval, localization, and comparison. Notationally, we say that an instance within the **Sher**lock corpus consists of an image i, a region specified by N bounding boxes $r = \{\langle x_{1i}, x_{2i}, y_{1i}, y_{2i} \rangle\}_{i=1}^{N}$, ¹⁰ a clue c corresponding to a literal description of r's contents, and an in Ference f that an annotator associated with i, r, and c. We consider:

¹⁰ As discussed in $\S3$, N has a mean/median of 1.17/1.0 across the corpus.

- J. Hessel et al.
- 1. Retrieval of Abductive Inferences: For a given image/region pair (i, r), how well can models select the ground-truth inference f from a large set of candidates (~1K) covering a broad swath of the corpus?
- 2. Localization of Evidence: Given an image *i* and an inference *f* written about an (unknown) region within the image, how well can models locate the proper region?
- 3. Comparison of Plausibility: Given an image/region pair (i, r) and a small set (~ 10) of relevant inferences, can models predict how humans will rank their plausibility?

Each task tests a complementary aspect of visual abductive reasoning (Fig. 4): retrieval tests across a broad range of inferences, localization tests within-images, and comparison tests for correlation with human judgement. Nonetheless, the same model can undertake all three tasks if it implements the following interface:

Sherlock Abductive Visual Reasoning Interface

- Input: An image i, a region r within i, and a candidate inference f.
- **Target:** A score s, where s is proportional to the plausibility that f could be inferred from (i, r).

That is, we assume a model $m : (i, r, f) \to \mathbb{R}$ that scores inference f's plausibility for (i, r). Notably, the interface takes as input inferences, but not clues: our intent is to focus evaluation on abductive reasoning, rather than the distinct setting of literal referring expressions.¹¹ Clues can be used for training m; as we will see in §5 our best performing model, in fact, does use clues at training time.

4.1 Retrieval of Abductive Inferences

For retrieval evaluation, at test time, we are given an (i, r) pair, and a large $(\sim 1 \text{K})^{12}$ set of candidate inferences $f \in F$, only one of which was written by an annotator for (i, r); the others are randomly sampled from the corpus. In the $im \rightarrow txt$ direction, we compute the mean rank of the true item (lower=better) and P@1 (higher=better); in the $txt \rightarrow im$ direction, we report mean rank (lower=better).

4.2 Localization of Evidence

Localization assesses a model's capacity select a regions within an image that most directly supports a given inference. Following prior work on literal referring expression localization [28,25,73] (inter alia), we experiment in two settings: 1) we are given all the ground-truth bounding boxes for an image, and 2) we are given only automatic bounding box proposals from an object detection model.

 $^{^{11}}$ In §B.1, for completeness, we give results on the retrieval and localization setups, but testing on clues instead.

¹² Our validation/test sets contain about 23K inferences. For efficiency we randomly split into 23 equal sized chunks of about 1K inferences, and report retrieval averaged over the resulting splits.

	Retrieval			Localization	Comparison	
	$\mathrm{im} \to \mathrm{txt}~(\downarrow)$	$\mathrm{txt} \to \mathrm{im}~(\downarrow)$	$P@1_{im \to txt}$ (\uparrow)	$\overline{\text{GT-Box}/\text{Auto-Box}}$ (\uparrow)	$\overline{\rm Val/Test~Human~Acc~(\uparrow)}$	
Random	495.4	495.4	0.1	30.0/7.9	1.1/-0.6	
Bbox Position/Size	257.5	262.7	1.3	57.3/18.8	5.5/1.4	
LXMERT	51.1	48.8	14.9	69.5/30.3	18.6/21.1	
UNITER Base	40.4	40.0	19.8	73.0/33.3	20.0/22.9	
CLIP ViT-B/16	19.9	21.6	30.6	85.3/38.6	20.1/21.3	
CLIP RN50x16	19.3	20.8	31.0	85.7/38.7	21.6/23.7	
CLIP RN50x64	19.3	19.7	31.8	86.6/39.5	25.1/26.0	
${\bf l}_{\rm s}$ + multitask clue learning	16.4	17.7	33.4	87.2/40.6	26.6/27.1	
Human + (Upper Bound)	-	-	-	92.3/(96.2)	42.3/42.3	

Table 2: Test results for all models across all three tasks. CLIP RN50x64 outperforms all models in all setups, but significant headroom exists, e.g., on Comparison between the model and human agreement.

GT bounding boxes. We assume an image i, the set of 3+ inferences F written for that image, and the (unaligned) set of regions R corresponding to F. The model must produce a one-to-one assignment of F to R in the context of i. In practice, we score all possible $F \times R$ pairs via the abductive visual reasoning interface, and then compute the maximum linear assignment [30] using lapjv's implementation of [24]. The evaluation metric is the accuracy of this assignment, averaged over all images. To quantify an upper bound, a human rater performed the assignment for 101 images, achieving an average accuracy of 92.3%.

Auto bounding boxes. We compute 100 bounding box proposals per image by applying Faster-RCNN [54] with a ResNeXt101 [69] backbone trained on Visual Genome to all the images in our corpus. Given an image i and an inference f that was written about the image, we score all 100 bounding box proposals independently and take the highest scoring one as the prediction. We count a prediction as correct if it has IoU > 0.5 with a true bounding box that corresponds to that inference,¹³ and incorrect otherwise.¹⁴

4.3 Comparison of Plausibility

We assess model capacity to make fine-grained assessments given a set of plausible inferences. For example, in Fig. 4c (depicting a group of men marching and carrying bags), human raters are *likely* to say that they are military men and that the photo was taken during WWII, and *unlikely* to see them as porters despite them carrying bags. Our evaluation assumes that a performant model's predictions should correlate with the (average) relative judgments made by humans, and we seek to construct a corpus that supports evaluation of such reasoning.

¹³ Since the annotators were able to specify multiple bounding boxes per OBSERVATION PAIR, we count a match to any of the labeled bounding boxes.

 $^{^{14}}$ A small number of images do not have a ResNeXt bounding box with IoU> 0.5 with any ground truth bounding box: in §5.1, we show that most instances (96.2%) are solvable with this setup.

Constructing sets of plausible inferences. We use a performant model checkpoint fine-tuned for the Sherlock tasks¹⁵ to compute the similarity score between all (i, r, f) triples in the validation/test sets. Next, we perform several filtering steps: 1) we only consider pairs where the negative inference received a higher score than the ground-truth according to the model; 2) we perform soft text deduplication to downsample inferences that are semantically similar; and 3) we perform hard text deduplication, only allowing inferences to appear verbatim 3x times. Then, through an iterative process, we uniquely sample a diverse set of 10 inferences per (i, r) that meet these filtering criteria. This results in a set of 10 plausible inference candidates for each of 485/472 validation/test images. More details are in §E. In a retrieval sense, these plausible inferences can be viewed as "hard negatives:" i.e., none are the gold annotated inference, but a strong model nonetheless rates them as plausible.

Human rating of plausible inferences. Using MTurk, we collect two annotations of each candidate inference on a three-point Likert scale ranging from 1 (bad: "irrelevant"/"verifiably incorrect") to 3 (good: "statement is probably true; the highlighted region supports it."). We collect 19K annotations in total (see §E for full details). Because abductive reasoning involves subjectivity and uncertainty, we expect some amount of intrinsic disagreement between raters.¹⁶ We measure model correlation with human judgments on this set via pairwise accuracy. For each image, for all pairs of candidates that are rated differently on the Likert scale, the model gets an accuracy point if it orders them consistently with the human rater's ordering. Ties are broken randomly but consistently across all models. For readability, we subtract the accuracy of a random model (50%) and multiply by two to form the final accuracy metric.

5 Methods and Experiments

Training objective. To support the interface described in §4, we train models $m : (i, r, f) \to \mathbb{R}$ that score inference f's plausibility for (i, r). We experiment with several different V+L backbones as detailed below; for each, we train by optimizing model parameters to score truly corresponding (i, r, f) triples more highly than negatively sampled (i, r, f_{fake}) triples.

LXMERT [61] is a vision+language transformer [64] model pre-trained on Visual Genome [29] and MSCOCO [33]. The model is composed of three transformer encoders [64]: an object-relationship encoder (which takes in ROI features+locations with a max of 36, following [2]), a language encoder that processes word tokens, and a cross modality encoder. To provide region information r, we calculate the ROI feature of r and always place it in the first object token to the visual encoder (this is a common practice for, e.g., the VCR dataset [75]).

¹⁵ Specifically, a CLIP RN50x16 checkpoint that achieves strong validation retrieval performance (comparable to the checkpoint of the reported test results in §5.1); model details in §5.

¹⁶ In §5.1, we show that models achieve significantly less correlation compared to human agreement.

We follow [9] to train the model in "image-text retrieval" mode by maximizing the margin m = .2 between the cosine similarity scores of positive triple (i, r, f)and two negative triples (i, r, f_{fake}) and (i_{fake}, r_{fake}, f) through triplet loss. **UNITER** [9] consists of a single, unified transformer that takes in image and text embeddings. We experiment with the Base version pre-trained on MSCOCO [33], Visual Genome [29], Conceptual Captions [57], and SBU Captions [41]. We apply the same strategy of region-of-reference-first passing and train with the same triplet loss following [9].

CLIP. We finetune the ViT-B/16, RN50x16, and RN50x50 versions of CLIP [51]. Text is represented via a 12-layer text transformer. For ViT-B/16, images are represented by a 12-layer vision transformer [10], whereas for RN50x16/RN50x64, images are represented by EfficientNet-scaled ResNet50 [16,62].

We modify CLIP to incorporate the bounding box as input. Inspired by a similar process from [76,70], to pass a region to CLIP, we simply draw a bounding box on an image in pixel space—we use a green-bordered / opaque purple box as depicted in Fig. 5b (early experiments proved this more effective than modifying CLIP's architecture). To enable CLIP to process the widescreen images of VCR, we apply it twice to the input using overlapping square regions, i.e., graphically, like this: $[1[2]_1]_2$, and average the resulting embeddings. We finetune using InfoNCE [59,40]. We sample a batch of truly corresponding (i, r, f)triples, render the regions r in their corresponding images, and then construct all possible negative (i, r, f_{fake}) triples in the batch by aligning each inference to each (i, r). We use the biggest minibatch size possible using 8 GPUs with 48GB of memory each: 64, 200, and 512 for RN50x64, RN50x16, and ViT-B/16, respectively.

Multitask learning. All models thus far only utilize inferences at training time. We experiment with a multitask learning setup using CLIP that additionally trains with clues. In addition to training using our abductive reasoning objective, i.e., InfoNCE on inferences, we mix in an additional referring expression objective, i.e., InfoNCE on clues. Evaluation remains the same: at test time, we do not assume access to clues. At training time, for each observation, half the time we sample an inference (to form (i, r, f), and half the time we sample a clue (to form (i, r, c)). The clue/inference mixed batch of examples is then handed to CLIP, and a gradient update is made with InfoNCE as usual. To enable to model to differentiate between clues/inferences, we prefix the texts with clue:/inference:, respectively.

Baselines. In addition to a random baseline, we consider a content-free version of our CLIP ViT-B/16 model that is given only the position/size of each bounding box. In place of the image, we pass a mean pixel value across the entire image and draw the bounding box on the image using an opaque pink box (see §5.2).

5.1 Results

Table 2 contains results for all the tasks: In all cases, our CLIP-based models perform best, with RN50x64 outperforming its smaller counterparts. Incorporating the multitask objective pushes performance further. While CLIP performs the

			the kitchen is part of	n a restaurant.	
	$P@1 (\uparrow) V$	$Val/Test$ Human (\uparrow)			
CLIP ViT-B/16	30.5	20.1/21.2	THE P		No Region
↓ Position only	1.3 18 1	5.5/1.4	ACTO DE		
No Context	24.8	18.1/17.8			
→ Only context ↓ Trained w/ only Clues	$18.9 \\ 23.0$	17.4/16.3 16.2/19.7			Only Context
└→ Crop no Widescreen	27.8	23.1/21.8			
Resize no Widescreen	27.7	19.4/20.6	Desitien Oak	Annual	
$\exists {\bf l} \exists {\bf k} \ {\rm Zero \ shot \ w/ \ prompt}$	12.0	10.0/9.5	Position Only	No Context	
(a)			(b)		

Fig. 5: We perform ablations by varying the input data, top (a), and the modeling components, bottom (a). Figure (b) depicts our image input ablations, which are conducted by drawing in pixel-space directly, following [76]. Having no context may make it difficult to situate the scene more broadly; here: neatly stacked cups could be in a bar, a hotel, a store, etc. Access only the context of the dining room is also insufficient. For modeling, bottom (a), cropping/resizing decreases performance on retrieval (P@1), but not comparison (Val/Test Human).

best, UNITER is more competitive on comparison and less competitive on retrieval and localization. We speculate this has to do with the nature of each task: retrieval requires models to reason about many incorrect examples, whereas, the inferences in the comparison task are usually relevant to the objects in the scene. In §C, we provide ablations that demonstrate CLIP models outperform UNITER even when trained with a smaller batch size. Compared to human agreement on comparison, our best model only gets 65% of the way there (27% vs. 42%).

5.2 Ablations

We perform data and model ablations on CLIP ViT-B/16. Results are in Fig. 5. Input ablations. Each part of our visual input is important. Aside from the position only model, the biggest drop-off in performance results from not passing the region as input to CLIP, e.g., P@1 for $im \rightarrow txt$ retrieval nearly halves, dropping from 31 to 18, suggesting that CLIP relies on the local region information to reason about the image. Removing the region's content ("Only Context") unsurprisingly hurts performance, but so does removing the surrounding context ("No Context"). That is, the model performs the best when it can reason about the clue and its full visual context jointly. On the text side, we trained a model with only clues; retrieval and comparison performance both drop, which suggests that clues and inferences carry different information (additional results in §B.1). Model ablations. We considered two alternate image processing configurations. Instead of doing two CLIP passes per image to facilitate widescreen processing (§5), we consider (i) center cropping and (ii) pad-and-resizing. Both take less computation, but provide less information to the model. Cropping removes the



Fig. 6: Validation retrieval Fig. 7: Error analysis: examples of false posperf. (P@1) vs. comparison itives and false negatives predicted by our acc. for CLIP checkpoints. model on the comparison task's validation set.

sides of images, whereas pad-and-resize lowers the resolution significantly. The bottom half of the table in Fig. 5a reports the results: both configurations lower performance on retrieval tasks, but there's less impact for comparison.

Better retrieval \rightarrow **better comparison.** In Fig. 6, we observe a high correlation between the retrieval performance of our (single-task) CLIP model checkpoints (P@1) and the comparison human accuracy for the comparison task. For the smaller RN50x16 and ViT-B/16 models, this effect cannot simply be explained by training time; for RN50x16, pearson corr. between training steps and comparison performance is 81, whereas, the correlation between P@1 and comparison performance is 91. Overall, it's plausible that a model with higher precision at retrieval could help further bridge the gap on the comparison task. Oracle text-only models are insufficient. One potential concern with our setup is that clues may map one-to-one onto inferences, e.g., if all soccer balls in our corpus were mapped onto "the owner plays soccer" (and vice versa). We compare to an oracle baseline that makes this pessimistic assumption (complementing our "No Context" ablation, which provides a comparable context-free visual reference to the clue). We give the model oracle access to the ground-truth clues. Following [6], we use T5-Large v1.1 [52] to map clues to inferences with no access to the image by fitting P(inference|clue) in a sequence-to-sequence fashion; training details are in §B. The resulting text-only clue \rightarrow inference model, when given the clue "chipped paint and rusted umbrella poles", estimates likely inferences, for example: "the area is in a disrepair", "the city does not care about its infrastructure.", etc. The text-only oracle under-performs vs. CLIP despite the fact that, unlike CLIP, it's given the ground-truth clue: on comparison, it achieves 22.8/19.3 val/test accuracy; significantly lower than 26.6/27.1 that our best vision+language model achieves. This is probably because global scene context cannot be fully summarized via a local referring expression. In the prior "chipped paint and rusted umbrella poles" example, the true inference, "this beach furniture does not get put inside at night", requires additional visual context beyond the clue-chipped paint and a rusty umbrella alone may not provide enough context to infer that this furniture is *beach* furniture.

5.3 Error Analysis

We conduct a quantitative error analysis of multitask CLIP RN50x64 for the comparison task. We select 340 validation images with highest human agreement, and split images into two groups: one where the model performed above average, and one where the model performed below average. We attempt to predict into which group an image will fall using logistic regression in 5-fold cross-validation. Overall, errors are difficult to predict. Surface level image/text features of the images/inferences are not very predictive of errors: relative to a 50% ROC AUC baseline, CLIP ViT-B/16 image features achieve 55%, whereas the mean SentenceBERT [53] embedding of the inference achieves 54%. While not available *a priori*, more predictive than content features of model errors are human Likert ratings: a single-feature mean human agreement model achieves 57% AUC, (more human agreement = better model performance).

Fig. 7 gives qualitative examples of false positives/negatives. The types of abductive reasoning the model falls short on are diverse. In the boat example, the model fails to notice that a florist has set up shop on a ship deck; in the window example, the model misinterprets the bars over the windows as being *outside* the building versus inside and attached to a bed-frame. The model is capable of reading some simple signs, but, as highlighted by [37], reasoning about the semantics of written text placed in images remains a challenge, e.g., a "no parking" sign is misidentified as an "okay to park" sign. Overall: the difficult-to-categorize nature of these examples suggests that the **Sherlock** corpus makes for difficult benchmark for visual abductive reasoning.

6 Conclusion

We introduce **Sherlock**, a corpus of visual abductive reasoning containing 363K clue/inference OBSERVATION PAIRS across 103K images. Our work complements existing abductive reasoning corpora, both in format (free-viewing, free-text) and in diversity (not human-centric). Our work not only provides a challenging vision+language benchmark, but also, we hope it can serve as a resource for studying visual abductive reasoning more broadly. Future work includes:

- 1. Salience: in **Sherlock**, annotators specify salient clues; how/why does salience differ from other free-viewing setups, like image captioning?
- 2. Ambiguity: when/why do people (justifiably) come to different conclusions?
- 3. Generative evaluation metrics: generation evaluation in abductive setting, i.e., without definitive notions of correctness, remains a challenge.

Acknowledgments. 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. AR was additionally in part supported by the DARPA PTG program, as well as BAIR's industrial alliance program. We additionally thank the UC Berkeley Semafor group for the helpful discussions and feedback.

References

- Aliseda, A.: The logic of abduction: an introduction. In: Springer Handbook of Model-Based Science, pp. 219–230 (2017)
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015)
- Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. TACL 6, 587–604 (2018)
- Berg, A.C., Berg, T.L., Daume, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., et al.: Understanding and predicting importance in images. In: CVPR (2012)
- Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., tau Yih, W., Choi, Y.: Abductive commonsense reasoning. In: ICLR (2020)
- Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR 3, 993–1022 (2003)
- Carson, D.: The abduction of sherlock holmes. International Journal of Police Science & Management 11(2), 193–202 (2009)
- Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: ECCV (2020)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
- 11. Du, L., Ding, X., Liu, T., Qin, B.: Learning event graph knowledge for abductive reasoning. In: ACL (2021)
- Fang, Z., Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Video2Commonsense: Generating commonsense descriptions to enrich video captioning. In: EMNLP (2020)
- Garcia, N., Otani, M., Chu, C., Nakashima, Y.: KnowIT vqa: Answering knowledge-based questions about videos. In: AAAI (2020)
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM (2021)
- 15. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
- Hobbs, J.R., Stickel, M.E., Appelt, D.E., Martin, P.: Interpretation as abduction. Artificial intelligence 63(1-2), 69–142 (1993)
- 18. Hosseini, H., Kannan, S., Zhang, B., Poovendran, R.: Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138 (2017)
- Ignat, O., Castro, S., Miao, H., Li, W., Mihalcea, R.: WhyAct: Identifying action reasons in lifestyle vlogs. In: EMNLP (2021)
- Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-QA: Toward spatio-temporal reasoning in visual question answering. In: CVPR (2017)
- Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)

- 16 J. Hessel et al.
- Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: CVPR (2016)
- Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)
- Jonker, R., Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38(4), 325–340 (1987)
- 25. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
- Kim, H., Zala, A., Bansal, M.: CoSIm: Commonsense reasoning for counterfactual scene imagination. In: NAACL (2022)
- Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
- Krahmer, E., Van Deemter, K.: Computational generation of referring expressions: A survey. Computational Linguistics 38(1), 173–218 (2012)
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2016)
- Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly 2(1-2), 83–97 (1955)
- Lei, J., Yu, L., Berg, T.L., Bansal, M.: TVQA+: Spatio-temporal grounding for video question answering. In: ACL (2020)
- 32. Lei, J., Yu, L., Berg, T.L., Bansal, M.: What is more likely to happen next? videoand-language future event prediction. In: EMNLP (2020)
- Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
- Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., Liu, J.: Violin: A largescale dataset for video-and-language inference. In: CVPR (2020)
- 35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
- Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
- Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: OCR-VQA: Visual question answering by reading text in images. In: ICDAR (2019)
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: FAccT (2019)
- 39. Niiniluoto, I.: Defending abduction. Philosophy of science 66, S436-S451 (1999)
- Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
- 41. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS (2011)
- 42. Ovchinnikova, E., Montazeri, N., Alexandrov, T., Hobbs, J.R., McCord, M.C., Mulkar-Mehta, R.: Abductive reasoning with a large knowledge base for discourse processing. In: IWCS (2011)
- 43. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. In: ICCV (2019)
- 44. Park, J.S., Bhagavatula, C., Mottaghi, R., Farhadi, A., Choi, Y.: VisualCOMET: Reasoning about the dynamic context of a still image. In: ECCV (2020)
- 45. Paul, D., Frank, A.: Generating hypothetical events for abductive inference. In: *SEM (2021)
- 46. Peirce, C.S.: Philosophical writings of Peirce, vol. 217. Courier Corporation (1955)

- 47. Peirce, C.S.: Pragmatism and pragmaticism, vol. 5. Belknap Press of Harvard University Press (1965)
- Pezzelle, S., Greco, C., Gandolfi, G., Gualdoni, E., Bernardi, R.: Be different to be better! a benchmark to leverage the complementarity of language and vision. In: Findings of EMNLP (2020)
- Pirsiavash, H., Vondrick, C., Torralba, A.: Inferring the why in images. Tech. rep. (2014)
- Qin, L., Shwartz, V., West, P., Bhagavatula, C., Hwang, J., Bras, R.L., Bosselut, A., Choi, Y.: Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In: EMNLP (2020)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
- Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bertnetworks. In: EMNLP (2019)
- 54. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. NeurIPS (2015)
- 55. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. In: ACL (2019)
- Shank, G.: The extraordinary ordinary powers of abductive reasoning. Theory & Psychology 8(6), 841–860 (1998)
- Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
- Shazeer, N., Stern, M.: Adafactor: Adaptive learning rates with sublinear memory cost. In: ICML (2018)
- 59. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NeurIPS (2016)
- 60. Tafjord, O., Mishra, B.D., Clark, P.: ProofWriter: Generating implications, proofs, and abductive statements over natural language. In: Findings of ACL (2021)
- Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: EMNLP (2019)
- Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
- Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: MovieQA: Understanding stories in movies through question-answering. In: CVPR (2016)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
- 65. Vedantam, R., Lin, X., Batra, T., Zitnick, C.L., Parikh, D.: Learning common sense through visual abstraction. In: ICCV (2015)
- Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A.: FVQA: Fact-based visual question answering. TPAMI 40(10), 2413–2427 (2017)
- 67. Wang, P., Wu, Q., Shen, C., Hengel, A.v.d., Dick, A.: Explicit knowledge-based reasoning for visual question answering. In: IJCAI (2017)
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P.,

Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: EMNLP: System Demonstrations (2020)

- 69. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
- Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: CPT: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
- 71. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: Collision events for video representation and reasoning. In: ICLR (2020)
- 72. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual Madlibs: Fill in the blank image generation and question answering. In: ICCV (2015)
- Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016)
- 74. Zadeh, A., Chan, M., Liang, P.P., Tong, E., Morency, L.P.: Social-iq: A question answering benchmark for artificial social intelligence. In: CVPR (2019)
- 75. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019)
- 76. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: MERLOT: multimodal neural script knowledge models. In: NeurIPS (2021)
- 77. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C.: Raven: A dataset for relational and analogical visual reasoning. In: CVPR (2019)
- Zhang, H., Huo, Y., Zhao, X., Song, Y., Roth, D.: Learning contextual causality from time-consecutive images. In: CVPR Workshops (2021)
- 79. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: Grounded question answering in images. In: CVPR (2016)