Pragmatic Issue-Sensitive Image Captioning

Allen Nie[1]   **Reuben Cohn-Gordon**[2]   **Christopher Potts**[1,2]

[1]Department of Computer Science   [2]Department of Linguistics
Stanford University
{anie, reubencg, cgpotts}@stanford.edu

Abstract

Image captioning systems need to produce texts that are not only *true* but also *relevant* in that they are properly aligned with the current issues. For instance, in a newspaper article about a sports event, a caption that not only identifies the player in a picture but also comments on their ethnicity could create unwanted reader reactions. To address this, we propose Issue-Sensitive Image Captioning (ISIC). In ISIC, the captioner is given a target image and an *issue*, which is a set of images partitioned in a way that specifies what information is relevant. For the sports article, we could construct a partition that places images into equivalence classes based on player position. To model this task, we use an extension of the Rational Speech Acts model. Our extension is built on top of state-of-the-art pretrained neural image captioners and explicitly uses image partitions to control caption generation. In both automatic and human evaluations, we show that these models generate captions that are descriptive and issue-sensitive. Finally, we show how ISIC can complement and enrich the related task of Visual Question Answering.

1 Introduction

Image captioning systems have improved dramatically over the last few years (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Hendricks et al., 2016; Rennie et al., 2017; Anderson et al., 2018), creating new opportunities to design systems that are not just accurate, but also produce descriptions that include relevant, characterizing aspects of their inputs. Many of these efforts are guided by the insight that high-quality captions are implicitly shaped by the communicative goal of identifying the target image up to some level of granularity (Vedantam et al., 2017; Mao et al., 2016; Luo et al., 2018; Cohn-Gordon et al., 2018).
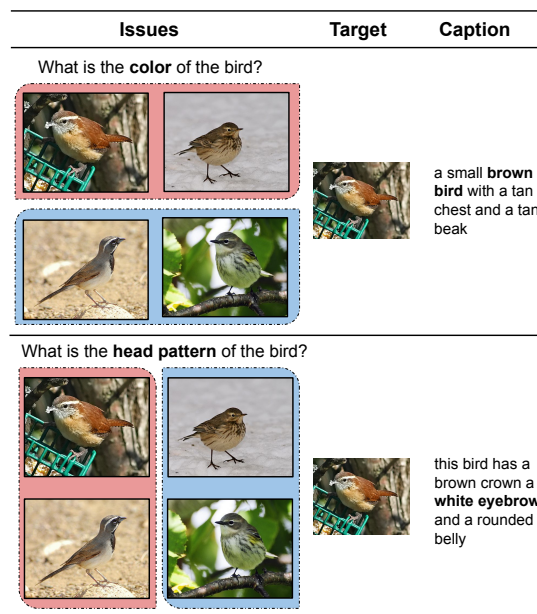
In this paper, we seek to more tightly control the

Figure 1: Examples highlighting the power of an issue-sensitive image captioner. Four images are partitioned in two ways, each capturing different issues by grouping them into equivalence classes. The first row contrasts the brown and grey color of the bird, and the second contrasts the existence of white eyebrows. The target image is the same in both cases, but the partition leads to different captions that key into the structure of the input issue.

information that a pretrained captioner includes in its output texts. Our focus is on generating captions that are *relevant* to the current issues. To see how important this can be, consider a newspaper article covering the action in a sports event. In this context, a caption that not only identified the player in a picture but also commented on their ethnicity could create unwanted reactions in readers, as it would convey to them that such information was somehow deemed relevant by the newspaper. On the other

hand, in an article about diversity in athletics, that same caption might seem entirely appropriate.

To push captioners to produce more relevant texts, we propose the task of Issue-Sensitive Image Captioning (ISIC). In ISIC, the captioner's inputs are image/issue pairs, where an *issue* is a set of images partitioned in a way that specifies what information is relevant. In our first example above, we might define a partition that grouped players into equivalence classes based on their team positions, abstracting away from other facts about them. For the second example, we might choose a more fine-grained partition based on position and demographic features. Given such inputs, the objective of the captioner is to produce a text that both accurately and uniquely describes the cell of the partition containing the target image. Figure 1 illustrates with examples from our own models and experiments.

In defining the task this way, we are inspired by Visual Question Answering (VQA; Antol et al. 2015), but ISIC differs from VQA in two crucial respects. First, we seek full image captions rather than direct answers. Second, our question inputs are not texts, but rather issues in the semantic sense: partitions on subsets of the available images. The ISIC module reasons about the cells in these partitions as alternatives to the target image, and our notion of relevance is defined in these terms. Nonetheless, VQA and ISIC complement each other: issues (as partitions) can be automatically derived from available image captioning and VQA datasets (Section 6), opening up new avenues for VQA as well.

Our models are built on top of pretrained image captioners with no need for additional training or fine-tuning. This is achieved by extending those models according to the Rational Speech Acts model (RSA; Frank and Goodman 2012; Goodman and Stuhlmüller 2013). RSA has been applied successfully to many NLP tasks (Section 2.3). Our key modeling innovation lies in building issues into these models. In this, we are inspired by linguistic work on question-sensitive RSA (Goodman and Lassiter, 2015; Hawkins and Goodman, 2019).

Our central experiments are with the Caltech-UC San Diego-Bird dataset (CUB; Welinder et al. 2010). This dataset contains extensive attribute annotations that allow us to study the effects of our models in precise ways. Using CUB, we provide quantitative evidence that our RSA-based models generate captions that both richly describe the tar-

get image and achieve the desired kinds of issue-sensitivity. We complement these automatic evaluation with a human evaluation in which participants judged our models to be significantly more issue-sensitive than standard image captioners. Finally, we show how to apply our methods to larger image captioning and VQA datasets that require more heuristic methods for defining issues. These experiments begin to suggest the potential value of issue-sensitivity in other domains that involve controllable text generation. We share code for reproducibility and future development at `https://github.com/windweller/Pragmatic-ISIC`.

2 Related Work

2.1 Neural Image Captioning

The task of image captioning crosses the usual boundary between computer vision and NLP; a good captioner needs to recognize coherent parts of the image and describe them in fluent text. Karpathy and Fei-Fei (2015) and Vinyals et al. (2015) showed that large-capacity neural networks can get traction on this difficult problem. Much subsequent work has built on this insight, focusing on two aspects. The first is improving image feature quality by using object-based features (Anderson et al., 2018). The second is improving text generation quality by adopting techniques from reinforcement learning to directly optimize for the evaluation metric (Rennie et al., 2017). Our work rests on these innovations – our base image captioning systems are those of Hendricks et al. (2016) and Rennie et al. (2017), which motivate and employ these central advancements.

There is existing work that proposes methods for controlling image caption generation with attributes. In general, these approaches involve models in which the attributes are part of the input, which requires a dataset with attributes collected beforehand. For instance, Mathews et al. (2015) collected a small dataset with sentiment annotation for each caption, Shuster et al. (2019) collected captions with personality traits, and Gan et al. (2017) with styles (such as humorous and romantic). The final metrics center around whether the human-generated caption was reproduced, or around other subjective ratings. By contrast, our method does not require an annotated dataset for training, and we measure the success of a model by whether it has resolved the issue under discussion.

1925

2.2 Visual Question Answering

In VQA, the model is given an image and a natural language question about that image, and the goal is to produce a natural language answer to the question that is true of the image (Antol et al., 2015; Goyal et al., 2017). This is a controllable form of (partial) image captioning. However, in its current form, VQA tends not to elicit linguistically complex texts; the majority of VQA answers are single words, and so VQA can often be cast as classification rather than sequence generation. Our goal, in contrast, is to produce linguistically complex, highly descriptive captions. Our task additionally differs from VQA in that it produces a caption in response to an *issue*, i.e., a partition of images, rather than a natural language question. In Section 4.2 and Section 6, we describe how VQA and ISIC can complement each other.

2.3 The Rational Speech Acts Model

The Rational Speech Acts model (RSA) was developed by Frank and Goodman (2012) with important precedents from Lewis (1969), Jäger (2007), Franke (2009), and Golland et al. (2010). RSA defines nested probabilistic speaker and listener agents that reason about each other in communication to enrich the basic semantics of their language. The model has been applied to a wide variety of diverse linguistic phenomena. Since RSA is a probabilistic model of communication, it is amenable for incorporation into many modern NLP architectures. A growing body of literature shows that adding RSA components to NLP architectures can help them to capture important aspects of context dependence in language, including referential description generation (Monroe and Potts, 2015; Andreas and Klein, 2016; Monroe et al., 2017), instruction following (Fried et al., 2018), collaborative problem solving (Tellex et al., 2014), and translation (Cohn-Gordon and Goodman, 2019).

Broadly speaking, there are two kinds of approaches to incorporating RSA into NLP systems. One class performs end-to-end learning of the RSA agents (Monroe and Potts, 2015; Mao et al., 2016; White et al., 2020). The other uses a pretrained system and applies RSA at the decoding stage (Andreas and Klein, 2016; Vedantam et al., 2017; Monroe et al., 2017; Fried et al., 2018). We adopt this second approach, as it highlights the ways in which one can imbue a wide range of existing systems with new capabilities.

2.4 Issue-Sensitivity in Language

Our extension of RSA centers on what we call *issues*. In this, we build on a long tradition of linguistic research on the ways in which language use is shaped by the issues (often called Questions Under Discussion) that the discourse participants regard as relevant (Groenendijk and Stokhof, 1984; Ginzburg, 1996; Roberts, 1996). Issues in this sense can be reconstructed in many ways. We follow Lewis (1988) and many others in casting an issue as a partition on a space of states into cells. Each cell represents a possible resolution of the issue. These ideas are brought into RSA by Goodman and Lassiter (2015) and Hawkins and Goodman (2019). We translate those ideas into the models for ISIC (Section 4), where an issue takes the form of a partition over a set of natural images.

3 Task Formulation

In standard image captioning, the input $\mathbf{i}$ is an image drawn from a set of images $\mathcal{I}$, and the output $\mathbf{w}$ is a sequence of tokens $[w_1, \ldots, w_n]$ such that each $w_i \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary.

In ISIC, we extend standard image captioning by redefining the inputs as pairs $(\mathbf{C}, \mathbf{i})$, where $\mathbf{C}$ is a partition[1] on a subset of elements of $\mathcal{I}$ and $\mathbf{i} \in \bigcup_{u \in \mathbf{C}}$. We refer to the partitions $\mathbf{C}$ as *issues*, for the reasons discussed in Section 2.4. The goal of ISIC is as follows: given input $(\mathbf{C}, \mathbf{i})$, produce a caption $\mathbf{w}$ that provides a true resolution of $\mathbf{C}$ for $\mathbf{i}$, which reduces to $\mathbf{w}$ identifying the cell of $\mathbf{C}$ that contains $\mathbf{i}$, as discussed in Section 2.4. Figure 2 presents an idealized example. (Figure 1 is a real example involving CUB and an ISIC captioner; see also Figure 3 and Figure 4 below.)

In principle, we could try to learn this kind of issue sensitivity directly from a dataset of examples $((\mathbf{C}, \mathbf{i}), \mathbf{w})$. We do think such dataset could be collected, as discussed briefly in Section 7. However, such datasets would be very large (each image needs to collect $|\mathbf{C}|$ number of captions), and our primary modeling goal is to show that such datasets need not be created. The issue-sensitive pragmatic model we introduce next can realize the goal of ISIC without training data of this kind.

---

[1]A set of sets $\mathbf{X}$ is a partition of a set X iff $u \cap v = \varnothing$ for all $u, v \in \mathbf{X}$ and $\bigcup_{u \in \mathbf{X}} = X$.
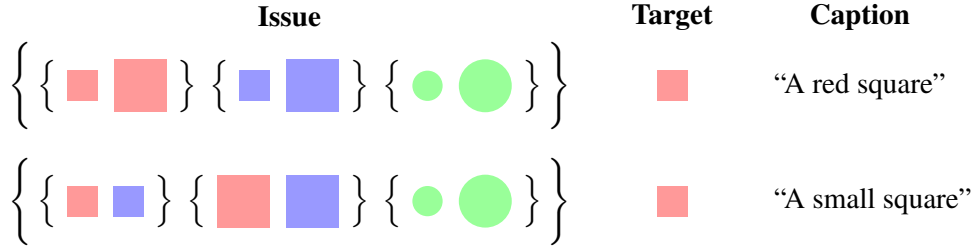
| | **Issue** | **Target** | **Caption** |

Figure 2: Two idealized examples highlighting the desired behavior for ISIC. A single set of images is partitioned in two ways. The top row groups them by color and shape, whereas the bottom row groups them by size and shape. A successful system for ISIC should key into these differences: for the same target image, its captions should reflect the partition structure and identify which cell the target belongs to, as in our examples. A caption like "A square" would be inferior in both contexts because it doesn't convey which cell the target image belongs to.

4 Models

4.1 Neural Pragmatic Agents

The models we employ for ISIC define a hierarchy of increasingly sophisticated speaker and listener agents, in ways that mirror ideas from Gricean pragmatics (Grice, 1975) about how meaning can arise when agents reason about each other in both production and comprehension (see also Lewis 1969).

Our base agent is a speaker $S_0(\mathbf{w} \mid \mathbf{i})$. In linguistic and psychological models, this agent is often defined by a hand-built semantics. In contrast, our S_0 is a trained neural image captioning system. As such, they are learned from data, with no need to hand-specify a semantic grammar or the like.

The pragmatic listener $L_1(\mathbf{i} \mid \mathbf{w})$ defines a distribution over states $\mathbf{i}$ given a message $\mathbf{w}$. The distribution is defined by applying Bayes' rule to the S_0 agent:

$$L_1(\mathbf{i} \mid \mathbf{w}) = \frac{S_0(\mathbf{w} \mid \mathbf{i})P(\mathbf{i})}{\sum_{\mathbf{i'} \in \mathcal{I}} S_0(\mathbf{w} \mid \mathbf{i'})P(\mathbf{i'})} \quad (1)$$

where $P(\mathbf{i})$ is a prior over states $\mathbf{i}$ (always flat in our work). This agent is pragmatic in the sense that it reasons about another agent, showing behaviors that align with the Gricean notion of *conversational implicature* (Goodman and Frank, 2016).

We can then define a pragmatic speaker using a utility function U_1, in turn defined in terms of L_1:

$$U_1(\mathbf{i}, \mathbf{w}) = \log L_1(\mathbf{i} \mid \mathbf{w}) \quad (2)$$

$$S_1(\mathbf{w} \mid \mathbf{i}) = \frac{\exp\left(\alpha U_1(\mathbf{i}, \mathbf{w}) - cost(\mathbf{w})\right)}{\sum_{\mathbf{w'}} \exp\left(\alpha U_1(\mathbf{i}, \mathbf{w'}) - cost(\mathbf{w'})\right)} \quad (3)$$

Here, α is a parameter defining how heavily S_1 is influenced by L_1. The term $cost(\mathbf{w})$ is a cost function on messages. In other work, this is often specified by hand to capture analysts' intuitions about complexity or markedness. In contrast, our version is entirely data-driven: we specify $cost(\mathbf{w})$ as $-\log(S_0(\mathbf{w} \mid \mathbf{i}))$.

4.2 Issue-Sensitive Speaker Agents

The agent in (3) has been widely explored and shown to deliver a powerful notion of context dependence (Andreas and Klein, 2016; Monroe et al., 2017). However, it is insensitive to the issues $\mathbf{C}$ that characterize ISIC. To make this connection, we extend (3) with a term for these issues:

$$U_1^{\mathbf{C}}(\mathbf{i}, \mathbf{w}, \mathbf{C}) = \log\left(\sum_{\mathbf{i'} \in \mathcal{I}} \delta_{[\mathbf{C}(\mathbf{i})=\mathbf{C}(\mathbf{i'})]} L_1(\mathbf{i'} \mid \mathbf{w})\right) \quad (4)$$

$$S_1^{\mathbf{C}}(\mathbf{w} \mid \mathbf{i}, \mathbf{C}) \propto \exp\left(\alpha U_1^{\mathbf{C}}(\mathbf{i}, \mathbf{w}, \mathbf{C}) - cost(\mathbf{w})\right) \quad (5)$$

where $\delta_{[\mathbf{C}(\mathbf{i})=\mathbf{C}(\mathbf{i'})]}$ is a partition function, returning 1 if $\mathbf{i}$ and $\mathbf{i'}$ are in the same cell in $\mathbf{C}$, else 0. This is based on a similar model of Kao et al. (2014). We use $\mathbf{C}(\mathbf{i})$ to denote the cell to which image $\mathbf{i}$ belongs under $\mathbf{C}$ (a slight abuse of notation, since $\mathbf{C}$ is a set of sets).

The construction of the partitions $\mathbf{C}$ is deliberately left open at this point. In some settings, the set of images $\mathcal{I}$ will have metadata that allows us to construct these directly. For example, in the CUB dataset, we can use the attributes to define intuitive partitions directly – e.g., the partition that groups images into equivalence classes based on the beak color of the birds they contain. The function can also be parameterized by a full VQA model $\mathcal{A}$. For a given question text $\mathbf{q}$ and image $\mathbf{i}$, $\mathcal{A}$ defines a map from $(\mathbf{q}, \mathbf{i})$ to answers $\mathbf{a}$, and so we can partition a subset of $\mathcal{I}$ based on equivalence classes defined by these answers $\mathbf{a}$.

4.3 Penalizing Misleading Captions

The agent in (5) is issue-sensitive in that it favors messages that resolve the issue $\mathbf{C}$. However, it does not include a pressure against hyper-specificity; rather, it just encodes the goal of identifying partition cells. This poses two potential problems.

The first can be illustrated using the top row of Figure 2. All else being equal, our agent (5) would treat "'A red square" and "A small red square" as equally good captions, even though the second includes information that is intuitively gratuitous given the issue. This might seem innocent here, but it can raise concerns in real environments, as we discussed in Section 1 in connection with our newspaper article examples.

The second problem relates to the data-driven nature of the systems we are developing: in being hyper-specific, we observed that they often mentioned properties not true of the target but rather only true of members of their equivalence classes. For example, in Figure 2, the target could get incorrectly described with "A large red square" because of the other member of its cell.

We propose to address both these issues with a second utility term U_2:

$$U_2(\mathbf{w}, \mathbf{i}, \mathbf{C}) = H(L_1(\mathbf{i'} \mid \mathbf{w}) \cdot \delta_{[\mathbf{C(i)}=\mathbf{C(i')}]}) \quad (6)$$

where H is the information-theoretic entropy. This encodes a pressure to choose utterances which result in the L_0 spreading probability mass as evenly as possible over the images in the target image cell. This discourages very specific descriptions of any particular image in the target cell, thereby solving both of the problems we identified above.

We refer to this agent as $S_1^{\mathbf{C}+H}$. Its full specification is as follows:

$$S_1^{\mathbf{C}+H}(\mathbf{w} \mid \mathbf{i}, \mathbf{C}) \propto$$
$$\exp\left(\alpha\left((1-\beta)U_1 + \beta U_2\right) - cost(\mathbf{w})\right) \quad (7)$$

where $\beta \in [0, 1]$ is a hyperparameter that allows us to weight these two utilities differently.

4.4 Reasoning about Alternative Captions

A pressing issue which arises when computing probabilities using (3), (5), and (7) is that the normalization constant includes a sum over all possible captions $\mathbf{w'}$. In the present setting, the set of possible captions is infinite (or at least exponentially large in the maximum caption length), making this computation intractable.

There are two solutions to this intractability proposed in the literature: one is to use S_0 to sample a small subset of captions from the full space, which then remains fixed throughout the computation (Andreas and Klein, 2016; Monroe et al., 2017). The drawback of this approach is that the diversity of captions that the S_1 can produce is restricted by the S_0. Since our goal is to generate captions which may vary considerably depending on the issue, this is a serious limitation.

The other approach is to alter the model so that the RSA reasoning takes place greedily during the generation of each successive word, word piece, or letter in the caption, so that the possible "utterances" at each step are drawn from a relatively small set of options to avoid exponential increase in search space (Cohn-Gordon et al., 2018). We opt for this *incremental* formulation and provide the full details on this model in Appendix A.

5 CUB Experiments

5.1 Preliminaries

Dataset The Caltech UC San Diego-Bird (CUB) dataset contains 11,788 images for 200 species of North American birds (Welinder et al., 2010). Each image contains a single bird and is annotated with fine-grained information about the visual appearance of that bird, using a system of 312 attributes (all of them binary) devised by ornithologists. The attributes have a *property::value* structure, as in *has_wing_color::brown*, and are arranged hierarchically from high-level descriptors (e.g., *bill*) to very specific low-level attributes (e.g., *belly pattern*). Appendix B provides a detailed example.

Reed et al. (2016) annotated each image in CUB with five captions. These captions were generated by crowdworkers who did not have access to the attribute annotations, and thus they vary widely in their alignment with the CUB annotations.

Constructing CUB Partitions CUB is ideal for testing our issue-sensitive captioning method because we can produce partitions directly from the attributes. For example, *has_wing_color::brown* induces a binary partition into birds with brown wings and birds with non-brown wings, and *has_wing_color* alone induces a partition that groups birds into equivalence classes based on their wing-color values. We selected the 17 most frequently appearing attributes, which creates 17 equivalence classes to serve as our issues.
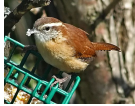
1928

Target	Issues		Issue-sensitive Caption	Base Caption
	Bill Shape		a small bird with a white breast and belly brown wings and tail and **a pointed beak**	this bird has a brown crown a white eyebrow and a brown and white breast
	Tail Pattern		a small brown and white bird with a long beak and **long tail feathers**	
	Belly Color		this is a bird with a **white belly** brown back and a brown head	this is a grey bird with a red head and a red beak
	Crown Color		this is a reddish orange bird with black and white wings and **a red crown**	

Figure 3: Captions for CUB. The left-hand cell in the 'Issues' column contains the target image, and right-hand cell is the union of the distractor cells. The 'Base Caption' texts are those produced by our S_0 model, and the 'Issue-sensitive Caption' texts were produced by $S_1^{\mathbf{C}+H}$.

Base Captioning System We trained a model released by Hendricks et al. (2016) with the same data-split scheme, where we have 4,000 images for training, 1,994 images for validation, and 5,794 images for testing. The model is a two-layer long short-term memory model (Hochreiter and Schmidhuber, 1997) with 1000-dimensional hidden size and 1000-dimensional word embeddings. We trained for 50 epochs with a batch size of 128 and learning rate $1e-3$. The final CIDEr score for our model is 0.52 on the test split. We use greedy decoding to generate our captions.

Feature-in-Text Classifier In order to examine the effectiveness of our issue-sensitive captioning models, we need to be able to identify whether the generated caption contains information regarding the issue. Even though each CUB image has a complete list of features for its bird, we must map these features to descriptions in informal text. For this, we require a text classifier. Unfortunately, it is not possible to train an effective classifier on the CUB dataset itself. As we noted above, the caption authors did not have access to the CUB attribute values, and so their captions tend to mention very different information than is encoded in those attributes. Furthermore, even if we did collect entirely new captions with proper attribute alignment, the extreme label imbalances in the data would remain a challenge for learning.

To remedy this, we use a sliding window text classifier. First, we identify keywords that can describe body parts (e.g. "head", "malar", "cheek-patch") and extract their positions in the text. Second, we look for keywords related to aspects (e.g., "striped", "speckled"); if these occur before a body-part word, we infer that they modify the body part. Thus, for example, if "scarlet and pink head" is in

the caption, then we infer that it resolves an issue about the color of the bird's head.

This classifier is an important assessment tool for us, so it needs to be independently validated. We meet this need using our human study in Section 5.4, which shows that our classifier is extremely accurate and, more importantly, not biased towards our issue-sensitive models.

5.2 Evaluating Attribute Coverage

We begin by assessing the extent to which our issue-sensitive pragmatic models produce captions that are more richly descriptive of the target image than a base neural captioner S_0 and its simple pragmatic variant S_1. For CUB, we can simply count how many attributes the caption specifies according to our feature-in-text classifier. More precisely, for each image and each model, we generate captions under all resolvable issues, concatenate those captions, and then use the feature-in-text classifier to obtain a list of attributes, which we can then compare to the ground truth for the image as given by the CUB dataset.

For S_0 and S_1, the captions do not vary by issue, whereas our expectation is that they do vary for $S_1^{\mathbf{C}}$ and $S_1^{\mathbf{C}+H}$. To further contextualize the performance of the issue-sensitive agents, we additionally define a model S_0 Avg that takes as inputs the average of all the features from all the images in the current partition, and otherwise works just like S_0. This introduces a rough form of issue-sensitivity, allowing us to quantify the value of the more refined approach defined by $S_1^{\mathbf{C}}$ and $S_1^{\mathbf{C}+H}$. Appendix D provides full details on how these models were optimized.

Table 1 reports on this evaluation. Precision for all models is very high; the underlying attributes in

	Precision	Recall	F1
S_0	**96.1**	16.6	28.3
S_0 Avg	94.0	32.7	48.5
S_1	93.9	29.6	45.0
$S_1^{\mathbf{C}}$	94.8	44.3	60.4
$S_1^{\mathbf{C}+H}$	94.7	**50.8**	**66.2**

Table 1: Attribute coverage results.

CUB are very comprehensive, so all high-quality captioners are likely to do well by this metric. In contrast, the recall scores vary substantially, and they clearly favor the issue-sensitive models, revealing them to be substantially more descriptive than S_0 and S_1. Figure 3 provides examples that highlight these contrasts: whereas the S_0 caption is descriptive, it simply doesn't include a number of attributes that we can successfully coax out of an issue-sensitive model by varying the issue.

The results also show that $S_1^{\mathbf{C}}$ and $S_1^{\mathbf{C}+H}$ provide value beyond simply averaging image features in **C**, as they both outperform S_0 Avg. However, it is noteworthy that even the rough notion of issue-sensitivity embodied by S_0 Avg seems beneficial.

Table 2 summarizes attribute coverage at the level of individual categories, for our four primary models. We see that the issue-sensitive models are clear winners. However, the entropy term in $S_1^{\mathbf{C}+H}$ seems to help for some categories but not others, suggesting underlying variation in the categories themselves.

5.3 Evaluating Issue Alignment

Our previous evaluation shows that varying the issue has a positive effect on the captions generated by our issue-sensitive models, but it does not assess whether these captions resolve individual issues in an intuitive way. We now report on an assessment that quantifies issue-sensitivity in this sense.

The question posed by this method is as follows: for a given issue **C**, does the produced caption precisely resolve **C**? We can divide this into two sub-questions. First, does the caption resolve **C**, which is a notion of *recall*. Second, does the caption avoid addressing issues that are distinct from **C**, which is a notion of *precision*. The recall pressure is arguably more important, but the precision one can be seen as assessing how often the caption avoids irrelevant and potentially distracting information, as discussed in Section 4.3.

Issues	S_0	S_1	$S_1^{\mathbf{C}}$	$S_1^{\mathbf{C}+H}$
wing pattern	3.9	22.1	**28.7**	14.9
belly pattern	7.6	19.4	32.0	**39.8**
breast pattern	7.6	17.0	30.8	**35.2**
nape color	6.0	18.1	31.2	**49.2**
upper tail color	1.3	24.5	**30.6**	22.6
under tail color	1.5	26.3	**33.1**	22.9
back color	5.1	24.9	32.6	**66.7**
leg color	6.2	**51.0**	45.6	7.4
throat color	16.4	46.2	66.0	**69.8**
crown color	49.5	50.4	77.4	**90.4**
bill shape	47.7	43.7	71.1	**91.8**
eye color	24.4	54.1	**61.3**	53.3
wing color	39.2	70.5	**82.7**	77.4
bill color	38.7	64.1	**80.4**	74.2
breast color	42.8	61.1	77.4	**90.4**
belly color	56.6	65.7	81.3	**93.3**
bill length	60.3	55.8	84.3	**95.5**

Table 2: F1 scores for each body part. The issue-sensitive models are superior for all categories except 'leg color'.

	Precision	Recall	F_1
S_0	10.5	21.1	15.5
S_0 Avg	12.1	29.0	17.0
S_1	11.2	21.7	14.8
$S_1^{\mathbf{C}}$	**18.7**	42.5	**25.9**
$S_1^{\mathbf{C}+H}$	16.6	**46.6**	24.5

Table 3: Issue alignment results.

Table 3 reports on this issue-sensitive evaluation, with F_1 giving the usual harmonic mean between our versions of precision and recall. Overall, the scores reveal that this is a very challenging problem, which traces to the fine-grained issues that CUB supports. Our $S_1^{\mathbf{C}+H}$ agent is nonetheless definitively the best, especially for recall.

5.4 Human Evaluation

We conducted a human evaluation of our models primarily to assess their issue sensitivity, but also to validate the classifier we used in the previous automatic evaluations.

Study Design We randomly sampled 110 images. For each, we have five conditions: our four RSA agents and the human captions from CUB. The items were arranged in a Latin Square design to ensure that no participant saw two captions for

Caption Source	Percentage	Size
S_0	20.9	273
S_1	24.5	273
$S_1^{\mathbf{C}}$	42.1	273
$S_1^{\mathbf{C}+H}$	**44.0**	273
Human	33.3	273

Table 4: Percentage of captions that contain an answer to the question according to our human evaluation. By Fisher's exact test, both issue-sensitive models ($S_1^{\mathbf{C}}$, $S_1^{\mathbf{C}+H}$) are different from issue-insensitive models (S_0, S_1) and humans ($p < 0.05$), but we do not have evidence for a difference between the issue-insensitive models.

Caption Source	Accuracy	Size
S_0	94.5	273
S_1	90.1	273
$S_1^{\mathbf{C}}$	92.7	273
$S_1^{\mathbf{C}+H}$	94.9	273
Human	90.5	273

Table 5: Classifier accuracy according to our human evaluation.

the same image and every participant saw an even distribution of conditions. We recruited 105 participants using Mechanical Turk. Each item was completed by exactly one participant. We received 1,365 responses in total.

Participants were presented with a question text and a caption and asked to use the caption to select an answer to the question or indicate that the caption did not provide an answer. (No images were shown, of course, to ensure that only the caption was used.) For additional details, see Appendix C.

Issue Sensitivity Table 4 shows the percentage of captions that participants were able to use to answer the questions posed. The pragmatic models are clearly superior. (The human captions are not upper-bounds, since they were not created relative to issues and so cannot vary by issue.)

Classifier Fairness We can also use our human evaluation to assess the fairness of the feature-in-text classifier (Section 5.1) that we used for our automatic evaluations. To do this, we say that the classifier is correct for an example x if it agrees with the human response for x. Table 5 presents these results. Not only are accuracy values very high, but they are similar for S_0 and $S_1^{\mathbf{C}+H}$.

6   MS COCO and VQA 2.0

The annotations in the CUB dataset allow us to generate nuanced issues that are tightly connected to the content of the images. It is rare to have this level of detail in an image dataset, so it is important to show that our method is applicable to less controlled, broader coverage datasets as well. As a first step in this direction, we now show how to apply our method using the VQA 2.0 dataset (Goyal et al., 2017), which extends MS COCO (Lin et al., 2014) with the question and answer annotations needed for VQA. While MS COCO does have instance-level annotations, they are mostly general category labels, so the attribute-dependent method we used for CUB isn't effective here. However, VQA offers a benefit: one can now control captions by generating issues from questions.

Dataset MS COCO contains 328k images that are annotated with instance-level information. The images are mostly everyday objects and scenes. A subset of them (204,721 examples) are annotated with whole image captions. Antol et al. (2015) built on this resource to create a VQA dataset, and Goyal et al. (2017) further extended that work to create VQA 2.0, which reduces certain linguistic biases that made aspects of the initial VQA task artificially easy. VQA 2.0 provides 1,105,904 question annotations for all the images from MS COCO.

Constructing Partitions To generate issues, we rely on the ground-truth questions and answers in the VQA 2.0 dataset. Here, each image is already mapped to a list of questions and corresponding answers. Given an MS COCO image and a VQA question, we identify all images associated with that question by exact string match and then partition these images into cells according to their ground-truth answers. Exactly the same procedure could be run using a trained VQA model rather than the ground-truth annotations in VQA 2.0.

Base Captioning System We use a pretrained state-of-the-art Transformer model with self-critical sequence training (Rennie et al., 2017). This has 6 Transformer layers with a 2048-dimensional hidden states, 512-dimensional input embeddings, and 8 attention heads at each layer. We use image features extracted by Anderson et al. (2018). The model achieves a CIDEr score of 1.29 for the test split. We use beam search (with beam size 5) to generate our captions.

Target	Question Text	Issues		Issue-sensitive Caption	Base Caption
	What position is this man playing?			**a pitcher** winding throwing ball on top of a field	a baseball player throwing a ball on a field
	What color is the wall?			a glass vase with **a red wall** with a chandelier	a vase with flowers in it on a table
	What color is the sky?			a **black and white** photo of an airplane in the sky	an airplane taking off from an airport runway
	How many toilets are there?			a bathroom with **two toilets** and a tub	a bathroom with a tub and a toilet and a window

Figure 4: Captions for MS COCO with issues determined by VQA 2.0. The left-hand cell in the 'Issues' column contains the target image, and right-hand cell is the union of the distractor cells. The 'Base Caption' texts are those produced by our S_0 model, and the 'Issue-sensitive Caption' texts were produced by $S_1^{\mathbf{C}+H}$.

Example Captions We show some examples for MS COCO in Figure 4. We chose these to highlight the potential of our model as well as remaining challenges. In datasets like this, the captioning model must reason about a large number of diverse issues, from objects and their attributes to more abstract concepts like types of food, sports positions, and relative distances ("How far can the man ride the bike?"; answer: "Far"). Our model does key into some abstract issues (e.g., "black and white photo" in row 2 of Figure 4), but more work needs to be done. Figure 4 also suggests shortcomings concerning over-informativity (e.g., the mention of a tub in response to an issue about toilets).

7 Conclusion

We defined the task of Issue-Sensitive Image Captioning (ISIC) and developed a Bayesian pragmatic model that allows us to address this task successfully using existing datasets and pretrained image captioning systems. We see two natural extensions of this approach that might be explored.

First, the method we proposed can be used as a method for assessing the quality of the underlying caption model. Using a dataset with issue annotations, if the model trained over the plain captions is more issue-sensitive, then it is better at decomposing the content of an image by its objects and abstract concepts.

Second, one could extend our notion of issue-sensitivity to other domains. As we saw in Section 6, questions (as texts) naturally give rise to issues in our sense where the domain is sufficiently structured, so these ideas might find applicability in the context of question answering and other areas

of controllable natural language generation.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 437–441, Minneapolis, Minnesota. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image

captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Michael Franke. 2009. *Signal to Act: Game Theory in Pragmatics*. ILLC Dissertation Series. Institute for Logic, Language and Computation, University of Amsterdam.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman, editor, *Language, Logic, and Computation*, volume 1, pages 221–237. CSLI, Stanford, CA.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA. Association for Computational Linguistics.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Noah D. Goodman and Daniel Lassiter. 2015. Probabilistic semantics and pragmatics. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2 edition. Wiley-Blackwell, Oxford.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 43–58. Academic Press, New York.

Jeroen Groenendijk and Martin Stokhof. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Robert X. D. Hawkins and Noah D. Goodman. 2019. Why do you ask? The informational dynamics of questions and answers. PsyArXiv, 6.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Gerhard Jäger. 2007. Game dynamics connects semantics and pragmatics. In Ahti-Veikko Pietarinen, editor, *Game Theory and Linguistic Meaning*, pages 89–102. Elsevier, Amsterdam.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

David Lewis. 1969. *Convention*. Harvard University Press, Cambridge, MA.

David Lewis. 1988. Relevant implication. *Theoria*, 54(3):161–174.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20. IEEE.

Alexander Mathews, Lexing Xie, and Xuming He. 2015. Senticap: Generating image descriptions with sentiments. *arXiv preprint arXiv:1510.01431*.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Will Monroe and Christopher Potts. 2015. Learning in the Rational Speech Acts model. In *Proceedings of 20th Amsterdam Colloquium*, Amsterdam. ILLC.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Craige Roberts. 1996. Information structure: Towards an integrated formal theory of pragmatics. In Jae Hak Yoon and Andreas Kathol, editors, *OSU Working Papers in Linguistics*, volume 49: Papers in Semantics, pages 91–136. The Ohio State University Department of Linguistics, Columbus, OH.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12516–12526.

Stefanie Tellex, Ross A. Knepper, Adrian Li, Thomas M. Howard, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems*.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the Cognitive Science Society*.

A Incremental Pragmatic Reasoning

The normalization terms of S_1, $S_1^{\mathbf{C}}$, and $S_1^{\mathbf{C}+H}$ all require a sum over all messages, rendering them intractable to compute.

We now describe a variant of the S_1 which performs pragmatic reasoning incrementally. This method extends in an obvious fashion to $S_1^{\mathbf{C}}$ and $S_1^{\mathbf{C}+H}$.

We begin by noting that a neural captioning model, at decoding time, generates a caption $\mathbf{w}$ one segment at a time (depending on the architecture, this segment may be a word, word piece, or character). We write $\mathbf{w} = (w_1 \ldots w_n)$, where w_i is the ith segment.

Concretely, a trained neural image captioner can be specified as a distribution over the subsequent segment given the image and previous words, which we write as $S_0(w_{n+1} \mid \mathbf{i}, [w_1 \cdots w_n])$. This allows us to define incremental versions of L_1 and S_1, as follows:

$$L_1(\mathbf{i} \mid w_{n+1}, [w_1 \ldots w_n]) \propto \\ S_0(w_{n+1} \mid \mathbf{i}, [w_1 \ldots w_n]) P(\mathbf{i}) \quad (8)$$

$$U_1(\mathbf{i}, w_{n+1}, [w_1, w_n]) = \\ \log L_1(\mathbf{i} \mid w_{n+1}, [w_1 \ldots w_n]) \quad (9)$$

$$S_1(w_{n+1} \mid \mathbf{i}, [w_1 \ldots w_n]) \propto \\ \exp(\alpha U_1(\mathbf{i}, w_{n+1}, [w_1, w_n]) - cost(w_{n+1})) \quad (10)$$

Here, we define the cost as the negative log-likelihood of the S_0 producing w_{n+1} given the image i and previous segments $[w_1 \ldots w_n]$. We can then obtain a caption-level model, which we term S_1^{INC} by contrast to the S_1 defined in (3):

$$S_1^{INC}(\mathbf{w} \mid \mathbf{i}) = \prod_{i=1}^{n} S_1(w_i \mid [w_1 \ldots w_{i-1}], \mathbf{i}) \quad (11)$$

$S_1^{INC}(\mathbf{w} \mid \mathbf{i})$ then serves as a tractable approximation of the caption-level $S_1(\mathbf{w} \mid \mathbf{i})$, and the same approach is easily extended to $S_1^{\mathbf{C}}$ and $S_1^{\mathbf{C}+H}$.

B   CUB Attribute Annotation Example

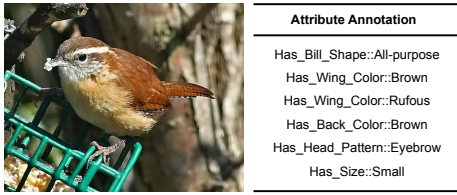| **Attribute Annotation** |
| --- |
| Has_Bill_Shape::All-purpose |
| Has_Wing_Color::Brown |
| Has_Wing_Color::Rufous |
| Has_Back_Color::Brown |
| Has_Head_Pattern::Eyebrow |
| Has_Size::Small |

Figure A1: A Carolina Wren from CUB. There can be multiple aspects per body part. Some general descriptors (e.g., *size*) do not have fine-grained aspects.

C   Human Study Design

Our study involved 110 randomly sampled images from CUB. For each, we have five conditions: original human caption, S_0 (base image caption model), S_1, $S_1^{\mathbf{C}}$, and $S_1^{\mathbf{C}+H}$. The items were arranged in a Latin Square design to ensure that no participant saw two captions for the same image and every participant saw an even distribution of conditions. We recruited 105 participants using Mechanical Turk. Each participant completed 13 items, and each item was completed by exactly one participant. We received a total of 1,365 responses. No participants or responses were excluded from our analyses.

In an instruction phase, participants were shown labeled images of birds to help with specialized terminology and concepts in the CUB domain. Before beginning the study, they were shown three examples of our task (Figure A2). These examples were chosen to familiarize participants with the underlying semantics of the captions. For example, if caption gives an ambiguous generic description like "the bird has a white body", it does not provide enough information to answer questions about the color of specific body parts, even though it does mention a color.

Following the example phase, we included a short trial phase of two example items, shown in Figure A3. We required participants to complete this trial before they started the study itself. We provided feedback immediately ("Wrong" or "'Correct') after they made selections in this phase.

Finally, an example item is given in Figure A4. Each annotation is structured in terms of a question (we rephrase issues in CUB as a question: *has_wing_color* is rephrased as *What is the wing color?*). Since the CUB attribute annotations consist of a *property::value* structure, we take the values associated with the property as our answer options.

D   Optimization Details

Computing infrastructure   Our experiment on CUB and MSCOCO is conducted on an NVIDIA TITAN X (Pascal) with 12196 MB graphic memory.

Computing time   Since we do not re-train our model, we report the inference time for our algorithm. Running our $S_1^{\mathbf{C}+H}$ on CUB test examples (5,794 images) takes about 40–50 minutes on a single GPU with specs listed above. For MS COCO, it takes substantially longer (about 1 minute per image) due to the more complex Transformer base image captioning model.

Hyperparameters   We did not conduct a hyperparameter search. We manually set hyperparameters for our RSA-based models. The hyperparameters include rationality α, entropy penalty β, and number of examples in a partition cell. The hyperparameters are chosen by small scale trial-and-error on validation data. We looked at the generated captions for 4 or 5 validation images of each model, and we decreased or increased our hyperparameters so that the generated captions for these images were coherent, grammatical, and issue-sensitive (when applicable). In Table A1, we report the hyperparameters we used for each RSA model.

Model	Cell Size	α	β
S_1	40	3	—
$S_1^{\mathbf{C}}$	40	10	—
$S_1^{\mathbf{C}+H}$	40	10	0.4

Table A1: Hyperparameter for each RSA model.

Validation performance   We report the performance on the validation set in Figure A2 and Figure A3, which contains 1,994 images. Even though the absolute numbers are slightly different from our result from the test set, the general trend still holds: issue-sensitive models significantly outperform issue-insensitive models.

	Precision	Recall	F1
S_0	**95.6**	16.1	27.5
S_1	93.6	27.8	42.9
$S_1^{\mathbf{C}}$	94.8	41.6	57.8
$S_1^{\mathbf{C}+H}$	94.4	**43.7**	**59.7**

Table A2: Attribute coverage results for validation set.

Figure A2: Example phase of the MTurk study

Figure A3: Trial phase of the MTurk study

Figure A4: An annotation item from the MTurk study. In this example, the correct answer is "curved (up or down)".

	Precision	Recall	F_1
S_0	10.8	28.1	15.6
S_1	11.0	20.8	14.4
$S_1^{\mathbf{C}}$	**18.5**	41.0	**25.5**
$S_1^{\mathbf{C}+H}$	14.8	**42.4**	22.0

Table A3: Issue alignment results for validation set.

E More Examples in CUB

We randomly sampled two test-set images (Figure A5 amd Figure A6) to show qualitatively how well our issue-sensitive caption model does compared to other models. To increase readability in these figures, we gloss issues as question texts.

Image ID: 121.Grasshopper Sparrow - Grasshopper Sparrow 0078 116052.jpg
Human caption: Small bird black tan and white feathers with a speck of yellow above beak.
S_0 **caption:** This bird has a white belly and breast with a speckled appearance on its back.

What is the eye/eyering/eyebrow color?
S_1 **caption:** A small tan and tan patterend bird with light belly.
$S_1^{\mathbf{C}}$ **caption:** This is a littl brown bird with black stripes on the crown and superciliary.
$S_1^{\mathbf{C}+H}$ **caption:** This is a bird with a white belly and a black and white spotted back.

What is the beak color?
S_1 **caption:** This bird has a white belly with dark spots spots and a short pointy beak.
$S_1^{\mathbf{C}}$ **caption:** This is a white and grey bird with an orange eyebrow and orange feet.
$S_1^{\mathbf{C}+H}$ **caption:** This is a small bird with a white belly and a brown back.

What is the breast color?
S_1 **caption:** A tan and tan sparrow connects to a white belly with tints of tan.
$S_1^{\mathbf{C}}$ **caption:** A small bird with a white belly and throat and a spotted brown back and head.
$S_1^{\mathbf{C}+H}$ **caption:** This bird has a white belly and breast with a brown crown and short pointy bill.

What is the belly color?
S_1 **caption:** A small light brown and white bird with dark eyes and a short red-tipped bill.
$S_1^{\mathbf{C}}$ **caption:** This bird has a white belly and breast with a speckled appearance elsewhere.
$S_1^{\mathbf{C}+H}$ **caption:** A small bird with a white belly and breast and a light brown crown and nape.

What is the beak length?
S_1 **caption:** A small round bird with multicolored tan and tan feathers.
$S_1^{\mathbf{C}}$ **caption:** This is a brown and tan speckled bird with a small beak and long tail feathers.
$S_1^{\mathbf{C}+H}$ **caption:** This bird has a speckled belly and breast with a short pointy bill.

Figure A5: Issue-sensitive and Issue-insensitive captions for an image of a Grasshopper Sparrow. To increase readability in these figures, we gloss issues as question texts.

Image ID: 030.Fish Crow - Fish Crow 0073 25977.jpg
Human caption: A large all black bird with a fluffy throat and thick round beak.
S_0 **caption:** This bird is completely black with a long blunt bill.

What is the crown color?
S_1 **caption:** An all black crow black shiny black bill legs feet and body.
$S_1^{\mathbf{C}}$ **caption:** This bird is all black with a long hooked bill and a think wingspan.
$S_1^{\mathbf{C}+H}$ **caption:** This is a mostly completely solid color bird with a black crown and a pointed bill.

What is the beak color?
S_1 **caption:** An all black crow black shiny black beak legs legs to shiny crow.
$S_1^{\mathbf{C}}$ **caption:** An all black crow with strong thick downward downward curved black beak and black legs.
$S_1^{\mathbf{C}+H}$ **caption:** This is a small pointy bird with a medium sized beak and is mostly black with a short beak.

What is the breast color?
S_1 **caption:** An all black crow black bill legs feet and body.
$S_1^{\mathbf{C}}$ **caption:** This stoutly black bird has strong legs and a long black beak.
$S_1^{\mathbf{C}+H}$ **caption:** This bird has a short black bill a white throat and a dark brown crown.

What is the belly color?
S_1 **caption:** An all jet black shiny black beak.
$S_1^{\mathbf{C}}$ **caption:** A solid black crow with strong claws and a trinagular jet-black belly.
$S_1^{\mathbf{C}+H}$ **caption:** This bird has a short bill a white belly and a a black crown.

Figure A6: Issue-sensitive and Issue-insensitive captions for an image of a Fish Crow. To increase readability in these figures, we gloss issues as question texts.