Pragmatic Inference with a CLIP Listener for Contrastive Captioning

Jiefu Ou[1]   **Benno Krojer**[2]   **Daniel Fried**[1]
Carnegie Mellon University[1]   Mila/McGill University[2]
`jiefuo@andrew.cmu.edu`

Abstract

We propose a simple yet effective and robust method for contrastive captioning: generating discriminative captions that distinguish target images from very similar alternative distractor images. Our approach is built on a pragmatic inference procedure that formulates captioning as a reference game between a *speaker*, which produces possible captions describing the target, and a *listener*, which selects the target given the caption. Unlike previous methods that derive both speaker and listener distributions from a single captioning model, we leverage an off-the-shelf CLIP model to parameterize the listener. Compared with captioner-only pragmatic models, our method benefits from rich vision-language alignment representations from CLIP when reasoning over distractors. Like previous methods for discriminative captioning, our method uses a hyperparameter to control the tradeoff between the informativity (how likely captions are to allow a human listener to discriminate the target image) and the fluency of the captions. However, we find that our method is substantially more robust to the value of this hyperparameter than past methods, which allows us to automatically optimize the captions for informativity — outperforming past methods for discriminative captioning by 11% to 15% accuracy in human evaluations.[1]

1 Introduction

Discriminative captioning provides a challenging testbed for generating context-sensitive grounded language. In this task, a model must produce a description of a *target image* (e.g., the green highlighted image in Figure 1) that allows a person to correctly identify the target image from among a set of similar *distractor images* (e.g., the red highlighted images). Good captions must strike a balance between two criteria: (1) being fluent

---

[1]The code is available at `https://github.com/JefferyO/prag_clip_contra_caption`
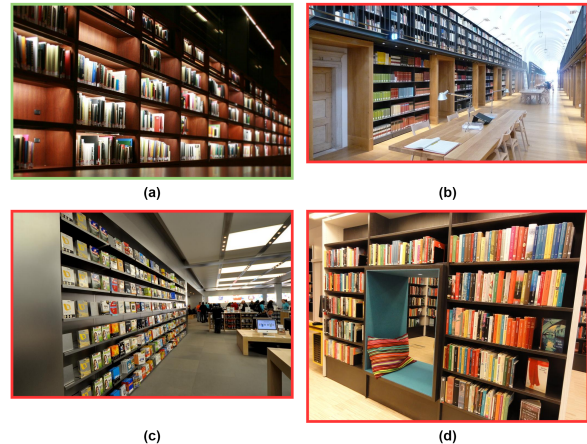
Base Speaker:	A bookshelf full of books in a library.
Incre-RSA:	A wooden bookshelf with many books on it.
E-S:	A wall of books on a wooden shelf.
PICL (Ours):	A wall of books in a library blurred by lights.
Human:	You can see several shelves of books that have lights on above the books. There are two thin banks of lights at the top right corner of the image.

Figure 1: Illustration of the contrastive captioning task with a random example from the ImageCoDe dataset. Models are tasked with generating captions that distinguish the target image (a) from other very similar distractors images (b) to (d). (There are a total of 9 distractors in each set of images, we omit the rest of them for simplicity of illustration.) Compared with baselines from previous work, our proposed approach, PICL, generates informative captions that help clearly identify the target out of the distractors, while remaining natural and fluent.

descriptions of the target image and (2) being discriminative in context: allowing a person to pick out the target image from the set.

Past work on discriminative captioning has successfully applied techniques from computational pragmatics to trade off between the two criteria above (Andreas and Klein, 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018). Possible captions are selected using a combination of two scoring functions: (1) the caption's probability under a standard image captioning model, or *base speaker score*, which measures the caption's fluency and

faithfulness to the image, and (2) a *base listener score*, which predicts how likely a human listener would be to correctly identify the target image given the caption, i.e. measuring discriminativeness. These past works typically obtain the listener scores from the image captioning (speaker) model itself, for example using Bayesian inference over the set of possible images (Cohn-Gordon et al., 2018). The relative weight of these two scores is controlled using a *informativity hyperparameter*,[2] whose value affects the tradeoff between producing captions that are predicted to be fluent and faithful, versus captions that are predicted to be discriminative. It is challenging to automatically choose a value for this hyperparameter, as captions that appear to be discriminative under a captioning model are frequently uninformative for people (Dessì et al., 2022).

Our approach, **PICL** (**P**ragmatic **I**nference with a **CLIP L**istener) follows this same pragmatic framework, but scores discriminativeness using a listener model separate from the speaker. We implement the listener model using CLIP (Radford et al., 2021). As shown in previous work, the rich vision-language representation learned in CLIP (1) provides robust assessments of model-generated captions that highly correlate with human judgments (Hessel et al., 2021), and (2) effectively quantifies the degree of discriminativeness/informativeness of visual referring expressions (Takmaz et al., 2022).

To evaluate PICL, we conduct experiments with sets of images from ImageCoDe (Krojer et al., 2022), a challenging dataset originally designed for contrastive retrieval: retrieving target images among a set of distractors given contextual descriptions. We perform contrastive captioning on this dataset for the first time. We compare PICL to past work on two criteria: (1) informativeness and (2) fluency, evaluating both metrics using automatic as well as human evaluations.

Results show that our approach typically outperforms past methods on both criteria, and is substantially more robust to the value of the informativity hyperparameter. In particular, we are able to choose this hyperparameter automatically by maximizing how informative the captions are predicted to be to human evaluators. In contrast, we find that maximizing predicted informativity leads past

---

methods to produce captions that are so disfluent that they are misleading for people. In this automatic hyperparameter selection setting, our method produces captions that are 11% to 15% easier for human annotators to interpret correctly than past work.

2 Related Work

Contrastive Captioning A variety of methods for contrastive captioning generate captions that optimize for discriminative objectives, e.g., minimizing the textual similarity between captions for the target and distractor images (Wang et al., 2020), using generated captions as input to image retrieval models (Luo et al., 2018; Liu et al., 2018), and computing CLIP similarity scores between captions and target images (Cho et al., 2022). Other methods involve leveraging fine-grained image regional features to generate distinctive captions based on similar and/or unique objects among target and distractors (Wang et al., 2021; Mao et al., 2022), paraphrasing generic captions to enhance both diversity and informativeness (Liu et al., 2019), and fine-tuning RL-optimized caption models to encourage low-frequency words (Honda et al., 2022). Most of the methods above require training a discriminative captioning model — either by designing an discriminative captioning architecture that takes multiple images as input, or fine-tuning a model using discriminative rewards. In contrast, our proposed approach is *fully inference-time* — it requires no training, and is applicable to any off-the-shelf generic captioning model.

Our approach builds on a family of inference-time pragmatic-based contrastive captioning methods which have taken one of two approaches: (1) incrementally generating captions but using only a captioning model (our *speaker* model), where tokens are chosen that have high probability for the target image and low probability for the distractor (Vedantam et al., 2017; Cohn-Gordon et al., 2018; Nie et al., 2020) or (2) using a separate discriminative model but selecting a discriminative caption from among a set of entire captions generated by the speaker model for the target image (Andreas and Klein, 2016; Luo and Shakhnarovich, 2017). Our work shows that these approaches can be productively combined, using a strong off-the-shelf discriminative model (CLIP) to guide the incremental generation of captions. This allows us to tackle a more challenging dataset and task than

previous discriminative captioning work, containing a large number (10) of highly-similar distractor images.

Pragmatics Our approach to contrastive generation follows a long line of work on computational pragmatics, particularly in the Rational Speech Acts framework (Frank and Goodman, 2012; Goodman and Frank, 2016) which models language generation as an interaction between speakers and listeners. Prior work has found that pragmatic generation can improve performance on a variety of NLP tasks, including reference games (Monroe et al., 2017), instruction generation (Fried et al., 2018), summarization (Shen et al., 2019), machine translation (Cohn-Gordon and Goodman, 2019), and dialogue (Kim et al., 2020; Fried et al., 2021).

Tradeoff between discriminativeness and accuracy/fluency Assessing the quality of image captions requires multifaceted evaluation. Prior work on contrastive/discriminative captioning investigates the tradeoff of model performance between discriminativeness and accuracy/fluency (Wang et al., 2021; Liu et al., 2019; Honda et al., 2022; Cho et al., 2022; Vedantam et al., 2017; Andreas and Klein, 2016). In this paper, we also perform an extensive study on the tradeoff between informativeness and fluency. Specifically, we focus on analyzing the robustness of the proposed and baseline methods in the tradeoff according to the selection of hyperparameters.

3 Method

Our PICL approach conducts incremental pragmatic inference at the token level by combining a base speaker and a CLIP listener to derive a pragmatic speaker. At each step of decoding, the base speaker selects a set of candidate tokens and adds them to partial captions. Given candidate partial captions, the listener updates its beliefs on which is the target among the set of images based on CLIP similarity measurement. In particular, it contrasts each partial caption to all the images by calculating the CLIP similarity scores of partial caption-image pairs and normalizes over all images to derive the listener likelihood. Finally, a pragmatic speaker reasons over both the base speaker and listener by combining their distribution to rerank partial captions, select a highly-scored subset and proceed to the next decoding step.

3.1 Incremental Pragmatic Inference Framework

Similar to Cohn-Gordon et al. (2018), we formulate the process of generating contrastive captions as a series of reference games between two agents, a *speaker* and a *listener*. Given a shared visual context $\mathcal{I} = i^+ \cup \mathcal{I}^-$ consisting of a target image i^+ and a set of m similar distractors $\mathcal{I}^- = \{i_1^-, \ldots, i_m^-\}$, the speaker aims to produce a sequence of T tokens $o_{1:T} = (o_1, \ldots o_T)$ that could let the listener identify i from I. Such pragmatic inference is conducted *incrementally*: at each step t of the caption generation, the speaker selects the next token o_t by playing the reference game with the listener based on the context I and the partial caption $o_{<t}$ obtained from the last step. In the following subsections, we will introduce the speaker and listener models as well as the incremental inference strategy in detail.

3.2 Speaker and Listener Models

Base Speaker At each step of generation, the *base speaker* S_0 yields a distribution $P_{S_0}(o_t|o_{<t}, i^+)$ over the token vocabulary for the next possible token o_t, conditioning on the previous partial caption and the target image. We parameterize P_{S_0} with a context-agnostic captioning model. In particular, we use OFA[3] (Wang et al., 2022), a unified sequence-to-sequence multimodal pretraining model and finetune it on MSCOCO Image Captioning dataset (Chen et al., 2015). Finetuned OFA is a strong base captioner; at the time of this work, it achieves state-of-the-art performance on MSCOCO Image Captioning.

Base Listener Given a candidate partial caption $o_{1:t} = (o_{<t}, o_t)$ generated by S_0, the base listener L_0 yields a distribution $P_{L_0}(i|o_{1:t}, \mathcal{I})$ over all candidate images $i \in \mathcal{I}$, modeling the likelihood of choosing each candidate given the partial caption at step t and the shared context $\mathcal{I}$. We derive P_{L_0} from a zero-shot CLIP model by normalizing its similarities between images and partial captions over all image candidates:

$$P_{L_0}(i|o_{1:t}, \mathcal{I}) = \frac{\exp(c(i, o_{1:t}))}{\Sigma_{i' \in \mathcal{I}} \exp(c(i', o_{1:t}))} \quad (1)$$

where $c(i, o_{1:t})$ denotes the cosine similarity between the CLIP visual encoding of i and textual

---

[3]We use the OFA-base configuration from `https://github.com/OFA-Sys/OFA`

encoding of $o_{1:t}$

Pragmatic Speaker From the base speaker and listener, we derive a distribution for the pragmatic speaker S_1 as

$$P_{S_1}(o_t|o_{<t}, i^+, \mathcal{I}) = P_{L_0}(i^+|o_{1:t}, \mathcal{I})^\lambda \\ \cdot P_{S_0}(o_t|o_{<t}, i^+)^{1-\lambda} \quad (2)$$

where $\lambda \in [0,1]$ is a "informativity" hyper-parameter that trades off between producing fluent (from S_0) and informative (from L_0) captions.

3.3 Decoding with Approximation

To iteratively generate captions with the pragmatic speaker S_1, we perform beam search with beam width B, which involves solving

$$\arg\max_{o_t} P_{S_1}(o_t|o_{<t}, i^+, \mathcal{I}) \quad (3)$$

for each beam item. However, it is computationally infeasible to obtain the exact solution to Equation 3 since it requires encoding all #(vocabulary size) possible next partial captions with CLIP to calculate P_{L_0} at each step. Thus, we adopt a sub-sampling approach similar to Andreas and Klein (2016); Fried et al. (2018). At each step of decoding, a subset of $N (N > B)$ candidate next partial captions $o_{1:T}$ are obtained via beam search from the base speaker distribution P_{S_0}, and these N candidates are rescored with Equation 2 to approximate Equation 3. Finally, only the top B candidates after rescoring are retained to continue with.

4 Experimental Setup

We evaluate PICL on ImageCoDe (Krojer et al., 2022), a dataset originally designed for image retrieval with contextual descriptions. Given the high visual similarity of the images in each problem in the dataset, we adopt it as a challenging testbed for discriminative captioning. We evaluate PICL and competitive baseline methods on two criteria, informativeness and fluency, using both automatic and human evaluation. For informativeness, we follow previous work (Cohn-Gordon et al., 2018; Newman et al., 2020) to automatically evaluate the performance of pragmatic models with an *evaluating listener* L_{eval}. The discriminativeness of the method being evaluated is quantified by the retrieval accuracy of L_{eval} with method-generated captions as input. For fluency, we score the well-formedness of

generated captions with the perplexity (PPL) under GPT-2 (Radford et al., 2019).

In addition to the automatic evaluation, we conduct human evaluation where annotators are tasked to a) retrieve the target image given the caption and b) score the fluency of the caption.

4.1 Dataset

We use sets of images collected in ImageCoDe (Krojer et al., 2022) to evaluate the proposed approach. Each image set in ImageCoDe consists of 10 visually similar images. The image sets are collected in two categories: *static pictures* and *video frames*. A random subset of images per set is selected as targets, for which human annotators write discriminative captions that are retained if other humans can successfully use it to retrieve the target.

In our experiments, we use the validation split of ImageCoDe for hyper-parameter selection and evaluate model performance on the test split. The valid and test sets contain 1,039 and 1,046 sets of images and 2,302 and 2,306 human written captions, respectively.

Table 1 shows the retrieval performance of several models on ImageCoDe test split, where **CLIP-zero-shot** is the base listener used in PICL and **ALBEF-finetuned** is the evaluating listener used for automatic evaluation (see Section 4.2). Given the large performance gap of all models between static and video subsets, we believe the video frames are too challenging for current neural models to make pragmatic and contextual inferences for both captioning and retrieving. Therefore, we use only static images in our experiments.

4.2 Automatic Evaluation

Informativeness Following Cohn-Gordon et al. (2018) and Newman et al. (2020), we evaluate the informativeness of captions generated by our method and baselines using a *listener test*: whether an *evaluative listener* model could identify the target image out of the distractors, given generated captions. However, an evaluative listener can only be an imperfect proxy for human listeners, and past work has found that utterances that are informative to an evaluative listener model can be uninterpretable to people, a phenomenon known as codebooking (Kim et al., 2019) or language drift (Lazaridou et al., 2020). This issue is particularly likely to complicate evaluation in a pragmatic framework like ours, where an explicit listener model (a frozen CLIP model, in our PICL

	All	Video	Static
CLIP-zero-shot	22.4	15.6	47.8
CLIP-finetuned-best	29.9	22.0	59.8
ALBEF-finetuned	33.6	22.7	74.2

Table 1: Retrieval accuracy on ImageCoDe test split with human-written contextual captions as input. In the proposed method, we use CLIP-zero-shot as the base listener and ALBEF-finetuned as the listener for evaluation. CLIP-finetuned denotes the best-performing model in previous work. The fine-tuned ALBEF outperforms the best CLIP model with a large margin on static images while improving slightly on video frames. Comparing with performances on static images, all models struggle on video frames.

approach) is used to guide utterance generation.

To mitigate this codebooking issue in evaluation, past work has made the evaluative listener dissimilar by training it on separate data (Cohn-Gordon et al., 2018; Kim et al., 2019; Fried et al., 2021); we additionally use a separate architecture for the evaluative listener, dissimilar from our CLIP listener: the ALBEF vision-language model (Li et al., 2020). We finetune ALBEF on the human-written contextual captions for the retrieval task in Image-CoDe.[4] As shown in Table 1, finetuned ALBEF outperforms the best-performing retrieval model from previous work (Krojer et al., 2022) on ImageCoDe with human-written captions, so we use ALBEF-finetuned as our evaluating listener in automatic evaluations of informativeness.

Fluency   While being informative, discriminative captions should also be natural and fluent. Therefore, we additionally perform automatic evaluations of the fluency of generated captions by computing their perplexity using a GPT-2 language model (Radford et al., 2019).

4.3   Human Evaluation

Recent analysis on ImageCode (Dessì et al., 2022) and in other reference game settings (Lazaridou et al., 2020) reveals that utterances generated by neural models can be discriminative enough for other neural models to retrieve the target image while being misleading to humans. This implies that the performance of a neural retriever evaluative listener (e.g., ALBEF) on model-generated captions might not correctly reflect the degree of informativeness of the captions from a human's

perspective. Therefore, we further conduct a human evaluation for PICL and baseline methods on Amazon MTurk, where we present human workers with the same image retrieval task as for ALBEF, and use the success rate of workers in identifying the correct target images (**retrieval accuracy**) to measure the informativeness of the given captions. To obtain human judgments of caption fluency, we additionally ask workers to score the captions on a Likert scale ranging from 1 (nonsense) to 5 (completely natural). We randomly sampled 100 sets of static images from the ImageCoDe test split and select one image with the human-written caption as the target. For each target, we produce a caption with each model and, together with the original human caption, present each caption-set pair to 3 workers. More details about the human evaluation setup could be found in Section A.3.

4.4   Baselines

We compare PICL to three baselines:

Base Speaker   We use the base speaker S_0 introduced in Section 3. The base speaker takes only the target image as input and generates context-agnostic captions regardless of the distractors.

Incre-RSA   We further implement the incremental RSA model (Incre-RSA) from Cohn-Gordon et al. (2018) as a competitive baseline. Specifically, we derive the Bayesian RSA model introduced in Cohn-Gordon et al. (2018) from our base speaker S_0, which enables direct comparison with our proposed approach. Unlike PICL, Incre-RSA does not have a separate model as the listener. The listener probabilities are derived with Bayesian inference at each decoding step based on the speaker distribution and an image prior.

E-S   Also based on S_0, we implement the *emitter-suppressor* (E-S) beam search introduced in Vedantam et al. (2017) for discriminative image captioning. Similar to Incre-RSA, the E-S approach differs from PICL mainly in that it does not contain a separate model to rescore partial captions from a listener's perspective. Instead, it incorporates contextual reasoning by selecting tokens that, under the base speaker, have high probability for the target image but low probability for the distractor images, using a weighted difference of scores. Since their task and model formulation considers only a single distractor image, we extend it to include all distractors in the set by calculating the suppressor distribution as the mean of the distribution of the

---

[4]Specifically, we finetuned the refcoco-checkpoint contrastively, i.e. with the 9 distractors in the same batch.
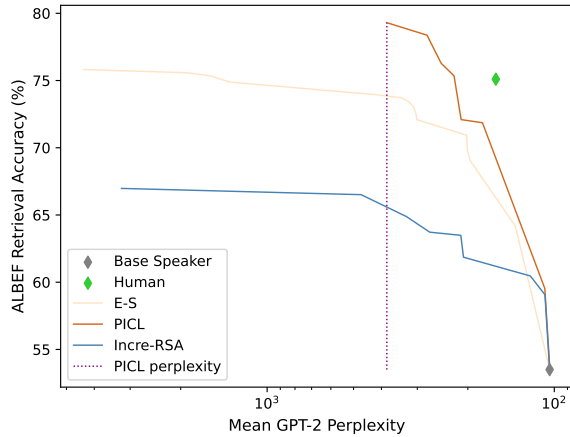
Figure 2: Automatic evaluations show a tradeoff between the informativeness (measured by ALBEF retrieval accuracy) and fluency (GPT-2 perplexity) of discriminative captions on the ImageCoDe valid set in automatic evaluations. Each curve is obtained by varying the value of the informativity hyperparameter. Compared with previous methods, our proposed PICL approach achieves a more robust trade-off between fluency and informativeness. The vertical line depicts the fluency-controlled criterion (Section 4.5), choosing a perplexity value that matches the perplexity of the maximally-informative PICL.

next token conditioned on each of the distractors.

For all three baselines, we use beam search at inference with the same beam width B as PICL.

4.5 Informativity Hyperparameter Selection

Both our PICL method and the Incre-RSA and E-S baselines use an informativity hyperparameter[5] to trade off between predicted informativity and fluency in generated captions. We describe two methods for choosing a value for this hyperparameter for each method.

Informativity Maximization In our primary set of experiments, we set the informativity hyperparameter for each method automatically to maximize the performance of our evaluating listener, ALBEF, on the captions in the validation set. We refer to the models obtained under this scheme as **PICL**, **Incre-RSA**, and **E-S**, respectively.

When maximizing predicted evaluative listener accuracy, we observe qualitatively that PICL typically generates captions which are fluent and human-understandable. In contrast, E-S and Incre-RSA are less robust, and under this informativity maximization objective typically produce highly

---

[5]Sometimes also referred to as a "rationality" parameter.

disfluent captions — identifying captions that are interpretable under our evaluating listener model, ALBEF, but potentially confusing to a human, consistent with past work identifying language drift in reference game setups (Lazaridou et al., 2020; Dessì et al., 2022). This trend is depicted in Figure 2, where optimizing for high ALBEF accuracy in E-S and Incre-RSA pushes the average GPT-2 perplexity of captions to extremely high values. We will see in human evaluations in Section 5 that the disfluent captions obtained by maximizing predicted informativity in the Incre-RSA and E-S baselines, though "understandable" to the ALBEF model, are often uninterpretable for humans.

Fluency Control Given the qualitative failures of E-S and Incre-RSA when maximizing automated proxies for informativity, we propose to improve these baselines using a fluency-controlled optimization scheme that pivots around PICL. In particular, we search for the informativity parameters for E-S and Incre-RSA so that the average GPT-2 perplexity of the generated captions are as close as possible to that of PICL. We refer to the models obtained under this scheme as **ES (PPL)** and **Incre-RSA (PPL)**.

5 Results

5.1 Automatic Evaluation

We use automatic evaluations (Section 4.2) to evaluate the tradeoff between the predicted informativity (using ALBEF) and predicted fluency (using GPT-2) of captions over a wide range of values for the informativity hyper-parameter of each method.

Hyper-parameter Sensitivity Figure 2 depicts how each method trades off between discriminativeness and fluency by varying the informativity hyper-parameter. PICL demonstrates higher robustness to hyper-parameter selection than Incre-RSA and ES in the trade-off: while optimizing for ALBEF-predicted informativity-maximization, Incre-RSA and ES produce more corrupted and disfluent captions with high perplexity whereas PICL's perplexity degrades less.

Informativeness As shown in Table 2, PICL substantially outperforms the base speaker and the incremental RSA (Incre-RSA, Cohn-Gordon et al. 2018) methods on ALBEF retrieval accuracy, and achieves comparable results to emitter-suppressor (E-S, Vedantam et al. 2017). The results demonstrate that our method could leverage CLIP as a

	ALBEF Accuracy	GPT-2 Perplexity
Human	74.2	138.4
Base Speaker	54.2	**99.4**
Optimized for Informativity		
Incre-RSA	64.3	2703.0
E-S	**77.5**	4093.6
PICL	77.3	380.2
Perplexity-Matched to PICL		
Incre-RSA (PPL)	62.9	446.5
E-S (PPL)	73.2	366.6

Table 2: Automatic evaluation results on the ImageCode test set: We evaluate informativity using the retrieval accuracy of the ALBEF evaluative listener using captions generated by each approach. PICL substantially outperforms Base Speaker, Incre-RSA, Incre-RSA (PPL), and E-S (PPL), achieving a competitive level of informativeness to E-S. In fluency, evaluated using GPT-2 perplexity, methods that control for the fluency (PPL) pivoting around PICL achieve similar level of perplexity, while E-S and Incre-RSA that optimized for informativity are substantially less fluent.

listener model in incremental pragmatic caption generation. For both E-S and Incre-RSA, controlling for fluency negatively affects ALBEF accuracy, which conforms with the trend in Figure 2.

Fluency Table 2 also shows the perplexity that GPT-2 assigns to the output of each model on the ImageCoDe test set. As discussed in Section 4.5, Incre-RSA and E-S are less robust when being optimized for informativity, which is reflected by their extremely high perplexity. In contrast, when controlling for the fluency to match PICL's validation perplexity, both Incre-RSA and E-S generate substantially more fluent captions with test perplexity similar to PICL, at the cost of predicted informativeness, as shown by a drop in ALBEF accuracy.

5.2 Human Assessment Performance

We perform human evaluations (Section 4.3) to validate these findings about the informativeness and fluency of the discriminative captioning methods.

Informativeness Human retrieval accuracies on model- and human-generated captions are depicted in Table 3. In the setting where models are automatically optimized for predicted informativity (Section 4.5), PICL substantially outperforms the Incre-RSA and E-S methods, with gains in human

Method	Human Accuracy	Fluency Rating
Human	**81.7**	4.76
Base Speaker	48.7	**4.80**
Optimized for Informativity		
Incre-RSA	50.7	2.87
E-S	54.0	3.59
PICL	**65.7**	4.07
Perplexity-Matched to PICL		
Incre-RSA (PPL)	53.3	4.23
E-S (PPL)	63.7	4.54

Table 3: Human evaluation results on 100 sets of images from ImageCoDe test split: Informativity is assessed by the retrieval accuracy of human annotators using captions generated by each approach. PICL outperforms all other models on human informativeness judgments. For fluency, human annotators evaluate using ratings on a 1-5 scale. Similar to results in the automatic evaluations of fluency (Table 2), annotators assign much lower fluency scores to E-S and Incre-RSA, which do not control for fluency.

accuracy of 11% and 15% respectively. The results indicate that captions generated by PICL are more informative than by other approaches, judged by human annotators. When we control the disfluency of the other methods to be similar to PICL (as measured by GPT-2 perplexity in automatic evaluations), PICL still substantially outperforms Incre-RSA (PPL) and slightly outperforms ES (PPL). Moreover, for both E-S and RSA, controlling for PPL results in more informative captions, which is not reflected in the automatic evaluations using ALBEF (Table 2), implying that disfluency has a more significant negative effect on informativity for humans. While past work has often relied only on automated evaluations, our results indicate that human evaluations are important to accurately compare the performance of discriminative captioning systems.

Fluency Table 3 also shows the average fluency scored by human workers for model- and human-generated captions. Similarly to Table 2 captions generated by E-S and Incre-RSA without controlling for perplexity are much more disfluent as scored by humans.

Informativity-Fluency Trade-off We further combine the human accuracy and fluency in Table 3 for each model and plot them in Figure 3.
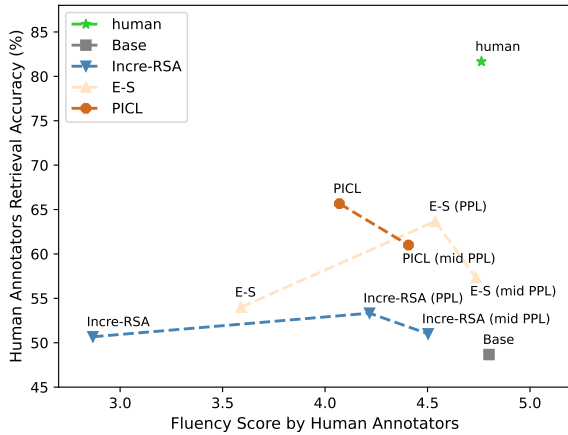
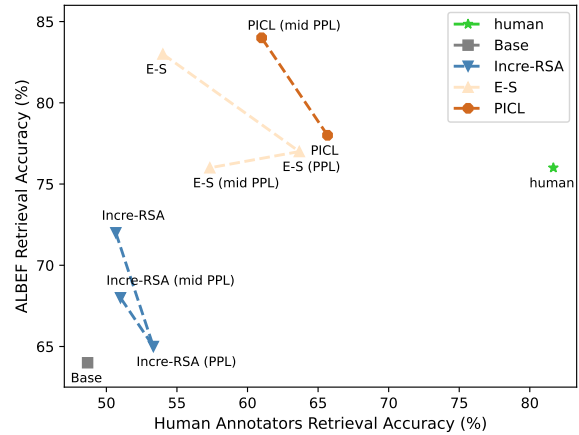Figure 3: Human eval results on 100 test split static image sets.

Figure 4: ALBEF accuracy and human accuracy are positively correlated for model-generated outputs, with the exception of disfluent captions produced by the variants of E-S and Incre-RSA that do not control for perplexity.

To depict the informativity-fluency trade-off under human assessments, we also include a setting of informativity hyperparameters for each method with an intermediate level of automatically predicted fluency. Specifically, for each model, we search for its informativity parameter so that the average GPT-2 perplexity of generated captions are as close as possible to the average perplexity of the base speaker + PICL. We refer to the models obtained under this scheme as **ES (mid PPL)**, **Incre-RSA (mid PPL)** and **PICL (mid PPL)**.

With the resulting plot shown in Figure 3, PICL outperforms Incre-RSA along both dimensions. In comparison with E-S, PICL achieves better discriminativeness with a loss in fluency. For E-S and Incre-RSA, the trade-off patterns are different from that under ALBEF (Figure 2). While optimizing for ALBEF accuracy consistently induces more disfluent generation, the optimal informativeness under human judgment is likely to be achieved with a moderate level of disfluency.

5.3 Automatic vs. Human Evaluation

The analysis above reflects both agreement and mismatch between automatic evaluation and human judgments on different aspects. To further reveal the correlation between them, and lay a foundation for future work on discriminative captioning to make automatic evaluations more predictive of human performance, we conduct analysis along both axes of informativity and fluency.

ALBEF vs. Human Retrieval Accuracy Figure 4 plots ALBEF against human retrieval accuracy on the same 100 sets of images. ALBEF accuracy has a strong positive correlation with human
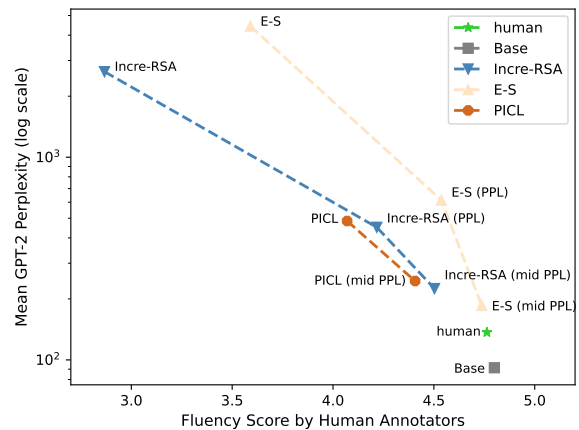

Figure 5: Mean perplexity, under GPT-2, is predictive of human fluency evaluations across systems.

judgments except for having human, E-S, and Incre-RSA as outliers. We posit that the performance mismatch on human written captions is because it is challenging for neural retrieval models like ALBEF to interpret human-written descriptions, which are highly nuanced and grammatically complex (Krojer et al., 2022). The high disfluency of the captions of E-S and and Incre-RSA hinders evaluators in interpreting them accurately, despite being discriminative to models.

GPT-2 Perplexity vs. Human Fluency Score As illustrated in Figure 5, on the 100 evaluation image sets, there is a strong correlation between the mean GPT-2 perplexity of captions and human fluency scores, implying that GPT-2 perplexity is a good proxy for human fluency judgments.

	Accuracy
PICL	77.3
- incremental	65.4
- distractors	57.5

Table 4: Automated ablation evaluations of informativeness. We evaluate "- incremental" that only conducts CLIP scoring and reranking on full captions generated by the speaker model, and "- distractor" in which only the target image is included during inference.

5.4 Ablation Results

To further understand the performance of PICL, we conduct ablation studies to investigate the role of 1) incremental pragmatic inference and 2) grounding language to distinguish from distractors.

For 1), we experiment with **PICL - incremental** that removes incremental inference by first using only the base speaker S_0 to generate a set of complete and context-agnostic captions, and using CLIP to score these entire captions. For 2), we evaluate **PICL - distractors**, excluding all distractors and providing only the target image during inference. At each decoding step, the listener distribution is derived by normalizing the CLIP similarities between partial captions and the target image over all candidates. As shown in Table 4, the retrieval accuracy drops substantially on both variations, suggesting that both the incremental inference and grounding to distractors are vital components for pragmatic reasoning in PICL.

6 Conclusion

We propose an incremental pragmatic inference approach with a CLIP listener, which combines the strengths of previous approaches that conduct incremental pragmatic reasoning with a separately modeled listener. We identify strengths and weaknesses of automatic model-based evaluation of discriminative captioning systems, and suggest that future work 1) control for the disfluency of generated captions and not solely optimize for predicted informativity and 2) use human evaluations. In human evaluations, our approach outperforms previous discriminative captioning methods, and is substantially more robust than previous approaches in trading off between the fluency and informativity of the captions to human listeners.

Limitations

We evaluate only on the "static" image partition of the ImageCoDe dataset. ImageCoDe contains another more challenging partition, containing frames from short temporal intervals in videos, which remains extremely difficult for all current discriminative captioning methods, including our PICL approach. (This partition, along with the static image partition that we use, has previously only been used in contrastive retrieval tasks, not in discriminative captioning.)

While we made a substantial effort to explore the tradeoff between informativity and fluency, we were limited in the number of human evaluations that we were able to do and could only evaluate a few settings of the informativity parameter for each method. We complement these human evaluations with automated evaluations on a much wider range of parameters, and analyze the correlations between human performance and judgements and the automated metrics.

References

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.

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

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, Seattle, United States. Association for Computational Linguistics.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss.

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.

Roberto Dessì, Eleonora Gualdoni, Francesca Franzon, Gemma Boleda, and Marco Baroni. 2022. Communication breakdown: On the low mutual intelligibility between human and neural captioning. In *Proceedings of EMNLP*.

Michael C Frank and Noah D Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

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.

Daniel Fried, Justin Chiu, and Dan Klein. 2021. Reference-centric models for grounded collaborative dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. 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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ukyo Honda, Taro Watanabe, and Yuji Matsumoto. 2022. Switching to discriminative image captioning by relieving a bottleneck of reinforcement learning. pages 1124–1134.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic Self-Consciousness.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Junnan Li, Yongkang Wong, Qi Zhao, and M. Kankanhalli. 2020. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22:554–565.

Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. 2019. Generating diverse and descriptive image captions using visual paraphrases. pages 4239–4248.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 353–369.

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

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111.

Yangjun Mao, Long Chen, Zhihong Jiang, Dong Zhang, Zhimeng Zhang, Jian Shao, and Jun Xiao. 2022. Rethinking the reference-based distinctive image captioning.

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.

Benjamin Newman, Reuben Cohn-Gordon, and Christopher Potts. 2020. Communication-based evaluation for natural language generation. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 116–126, New York, New York. Association for Computational Linguistics.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. In *Proceedings of NAACL*.

Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.

Ramakrishna Vedantam, Samy Bengio, Kevin P. Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B. Chan. 2020. Compare and reweight: Distinctive image captioning using similar images sets. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 370–386.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B. Chan. 2021. Group-based distinctive image captioning with memory attention. In *Proceedings of the 29th ACM International Conference on Multimedia*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.

A Implementation Details

A.1 Computational Resources

The finetuning of OFA model on COCO captions is run on $4 \times$ Tesla V100 32GB GPUs.

All pragmatic inference experiments are run on $4 \times$ GeForce RTX 2080 Ti GPUs.

A.2 Hyperparameter Searching

A.2.1 Rationality Parameters

Searching Range The search ranges of the rationality parameter for PICL, E-S, and Incre-RSA are [0, 1], [0, 1], [0, 2] respectively.

Searching Method We conduct all the hyperparameter searching via coarse-to-fine search, with step sizes 0.1, 0.01, and 0.001 respectively.

A.2.2 Beam Search Parameters

For beam search parameters B, N discussed in Section 3.2, we set $B = 16$ and $N = 256$.

A.3 Human Evaluation

Figure 6 shows an example interface of the human evaluation. We have three MTurk workers evaluate each of the 100 instances of (images, caption) for each of the ten configurations of methods (including human-written captions) for informativity (by requiring them to choose the image referred to by the caption) and fluency (on a 1-5 Likert scale) . Workers are paid with $0.15 per caption evaluation.

Instructions:

Given the description and the set of 10 images below,

1. Select the image that is best described by the description.

2. Score the fluency of the description.

Click on the images to display them in full size.

Description:

White table and chairs behind bright green plants.

Images:

1	2	3	4	5

6	7	8	9	10

1. Which image is best described by the description?

Select an option   (Please select) ⌄

2. How fluent is the description?

○ 5: Fluent
○ 4
○ 3: Slightly ungramatical or unnatural, but understandable
○ 2
○ 1: Totally ungramatical or unnatural

Submit

Figure 6: Human Evaluation Interface

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section, after Section 6

☒ A2. Did you discuss any potential risks of your work?
No, we do not foresee potential risks of this work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract; Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.1

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The artifact is publicly available under the MIT license

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data we used is publicly available

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1

C ☑ Did you run computational experiments?

Section 4.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.2

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
All the experiments are done in single runs

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.2

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Section 4.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.3

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A.3

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We mention the data usage if for research purposes and could provide the emails to the workers if needed

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We collect data using Amazon Mechanical Turk

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We could provide it later if needed