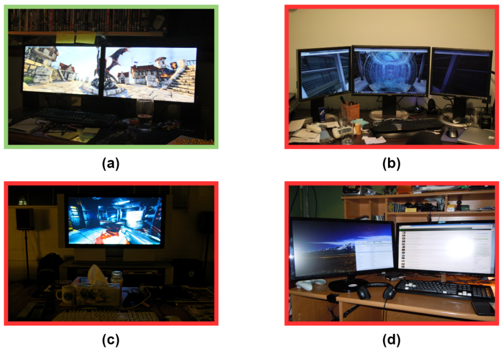
Incremental Pragmatic Inference with a CLIP Listener for Contrastive Captioning

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 efficient method for contrastive captioning: generating discriminative captions that distinguish target images from very similar alternative distractor images. Our approach is built on an incremental pragmatics inference procedure that formulates captioning as a series of reference games between a speaker and a listener, one at each token generation step in decoding. Unlike previous methods that derive both speaker and listener distributions from a single captioning model, we leverage an off-the-shelf zero-shot 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. On a challenging dataset, we demonstrate the proposed method better balances the trade-off between discriminativeness and fluency: compared with the two competitive prior approaches, our method either generates significantly more informative [1] captions with moderate loss in fluency or shows a comparable level of descriptiveness with much more natural utterances.



Base Speaker:	A computer monitor on a desk with a glass of beer
Incre-RSA:	A computer monitor with a screenshot on it
E-S:	A soda glass a television and a book shelf
PICL (Ours):	A gaming screen that has a medieval scene on it on a table
Human:	Two computer screens have split a dragon right down the middle.

Figure 1: Illustration of the contrastive captioning task with a random example from the ImgeCoDe 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 previous methods, our proposed approach could generate informative captions that help clearly identify the target out of the distractors, while being natural and fluent.

1 Introduction

Successful communication requires using language *pragmatically*: producing utterances that are contextually appropriate. Modeling the pragmatic aspect of language is thus an important step for developing computational models that communicate, interact, and cooperate with human users. In this paper, we model the pragmatics of language grounding in a visual context. In particular, we focus on contrastive (discriminative) image captioning: producing captions that are not only literally true but also distinguish target images from similar distractors, as illustrated in Figure 1.

We introduce an inference-time method for contrastive captioning that performs pragmatic reasoning incrementally: at each step of caption generation, the process of yielding the next token is formulated as a reference game between a speaker and a listener. The speaker is tasked to produce the next token that updates the partial caption in a way to help the listener identify the target image. And the listener aims at picking the correct target described by the speaker-provided partial caption out of a set of distractors. To generate a caption that is informative in the game context, a *pragmatic* speaker will select the next token by inferring how a listener will retrieve the target image out of distractors given different partial caption candidates.

Compared with previous work that derives prag-

[1]We use discriminativeness and informativeness interchangeably to describe how informative a reference expression (caption) is to distinguish a target image from distractors.

matic speakers and listeners from only a captioner model via Bayesian inference (Cohn-Gordon et al., 2018) or modified beam search objective (Vedantam et al., 2017), we model the behaviour of the listener using CLIP (Radford et al., 2021). As shown in previous work, the rich vision-language representation learned in CLIP could 1) provide robust assessments of model-generated captions that highly correlate with human judgments (Hessel et al., 2021), and 2) effectively quantify the degree of discriminativeness/informativeness of visual referring expressions (Takmaz et al., 2022). Our method leverages these properties to guide discriminative caption generation. To this end, we propose **PICL**, a method of **P**ragmatic **I**nference with a **CLIP** **L**istener for contrastive captioning. While CLIP is pretrained on images with *full* text descriptions, we find that we are nevertheless able to integrate CLIP in the scoring of *partial* captions, finding that incremental rescoring is superior to choosing from a set of full captions.

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 from distractors given contextual descriptions. Our method allows us to perform contrastive captioning on this dataset for the first time.

We automatically evaluate proposed methods and baselines from two aspects: *informativeness* and *fluency*. For informativeness evaluation, we obtain a competitive retrieval model by fine-tuning a large-scale vision-language pretraining model, ALBEF (Li et al., 2021a), on ImageCoDe to achieve state-of-the-art retrieval accuracy on human-written captions. We then use this model's retrieval accuracy on method-generated captions to measure how discriminative the captions are. To evaluate fluency, we score the perplexity of generated captions with an off-the-shelf language model (Radford et al., 2019).

Results show that our approach achieves competitive or better performance to past work on both axes: compared with previous methods which formulate pragmatic inference using just a captioning model, PICL significantly outperforms one of them (Cohn-Gordon et al., 2018) on informativeness with only a slight drop in fluency and achieves competitive discriminative performance to the other (Vedantam et al., 2017) with significantly more fluent generation (as shown in the example in Figure 1).

2 Related Work

Contrastive Captioning Prior work on contrastive captioning has 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).

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[2] (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 encoding of $o_{1:t}$

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

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

where $\lambda \in [0, 1]$ is a "rationality" hyper-parameter that trades off between producing context-agnostic (from S_0) and discriminative (from L_0) language.

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 subsampling 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. Following previous work (Cohn-Gordon et al., 2018; Newman et al., 2020), we automatically evaluate the performance of pragmatic models with an *evaluating listener* L_{eval}. The discriminativeness of the method being evaluated is quantified by the accuracy of L_{eval} identifying correct target images with method-generated captions as input.

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

4.1 Dataset

We use sets of images collected in ImageCoDe to evaluate the proposed approach. Each image set in ImgaeCoDe consists of 10 visually similar images. The image sets are collected in two categories: *static pictures* and *video frames*. Each static picture set is constructed by nearest neighbor retrieval in the Open Images dataset (Kuznetsova et al., 2020) using the CLIP visual encodings; each video frame set is collected by sampling frames from the same scene from various video datasets (Li et al., 2020; Xu et al., 2016; Das et al., 2013). A random subset of images per set is selected as targets, for which human annotators write discriminative captions.

In our experiments, we use the validation split of ImageCoDe for hyper-parameter selection and evaluate model performance on the test split. During the evaluation, every model generates a caption for each target image that has a human-written caption (for comparison). The valid and test sets contain 1,039 and 1,046 sets of images and on average 2.22 and 2.20 target images with human written captions per set, respectively.

4.2 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. 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 next token conditioned on each of the distractors. 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 suppressing tokens that are likely in the speaker distribution.

For all three baselines, we use beam search at inference with the same beam width B as PICL (subsection 3.3).

4.3 Automatic Evaluation

	all	video	static
CLIP-zero-shot	22.4	15.6	47.8
CLIP-finetuned-best	29.9	22.0	59.8
ALBEF-fientuned	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

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 a trained listener model could identify the target out of the distractors given the generated captions. We develop the listener for evaluation (L_{eval}) with ALBEF (Li et al., 2021b), another vision-language pretraining model that learns to align image-text representation before fusing them through cross-modal attention. In adaption to ImageCoDe data, ALBEF is finetuned with human-written contextual captions for the retrieval task. As shown in Table 1, finetuned ALBEF outperforms the previous best-performing retrieval model on ImageCoDe with human-written captions, demonstrating its effectiveness as a high-performing listener for evaluating discriminative captions.

Fluency While being informative, a desired discriminative caption should also be natural and well-formed. Therefore, we additionally measure the fluency of generated captions by scoring them using a language model. Specifically, we calculate the perplexity of each caption with GPT-2 (Radford et al., 2019).

	all	video	static
Human	33.6	22.7	74.2
Base Speaker	27.9	20.9	54.2
Incre-RSA	32.8	25.4	64.7
E-S	**38.5**	**27.7**	**79.0**
PICL	38.1	**27.7**	77.3

Table 2: Informativeness evaluation: The retrieval accuracy of ALBEF using captions generated by each approach on the ImageCoDe test set. PICL substantially outperforms Base Speaker and Incre-RSA, achieving a competitive level of informativeness to E-S. Captions generated by both E-S and PICL are more discriminative (as measured by ALBEF) than human-written descriptions.

	all	video	static
Human	105.1	96.2	138.4
Base Speaker	**91.6**	**89.5**	**99.4**
Incre-RSA	206.1	176.8	315.9
E-S	587.6	519.4	844.0
PICL	246.6	211.0	380.2

Table 3: Fluency evaluation: GPT-2 perplexity of model- and human-generated captions on the ImageCoDe test split. The captions from PICL are slightly less natural and fluent compared with those of Incre-RSA, while being substantially better than E-S's captions.

5 Results and Analysis

Informativeness As shown in Table 2, the proposed PICL approach substantially outperforms the base speaker and the incremental RSA inference on the ALBEF retrieval accuracy and achieves comparable results to the emitter-suppressor beam search. The results demonstrate that our method could effectively leverage CLIP as a listener model in incremental pragmatic caption generation. Moreover, finetuned ALBEF attains higher retrieval accuracy with both ES and PICL's output than with human captions. While it indicates captions from ES and PICL are equally or more informative than human captions to ALBEF, whether these captions are also similarly discriminative for humans to retrieve targets still needs to be validated through human evaluations in future work.

Fluency Table 3 shows the perplexity that GPT-2 assigns to the output of each model. In combination with Table 2, it demonstrates that the proposed approach achieves a superior trade-off between discriminativeness and fluency: it outperforms Incre-RSA by a large margin on informativeness while sacrificing a moderate drop in fluency, and it achieves comparable discriminative performance to E-S with the generated captions significantly more fluent to the language model. However, all of the captions from pragmatic models are far less fluent than human captions, indicating that models still fall behind humans in generating captions that are both informative and natural. In Figure 2, we illustrate the performance on fluency versus perplexity of different models on the valid set of ImageCode. It confirms that compared with previous methods, PICL better balances discriminativeness and fluency trade-off.

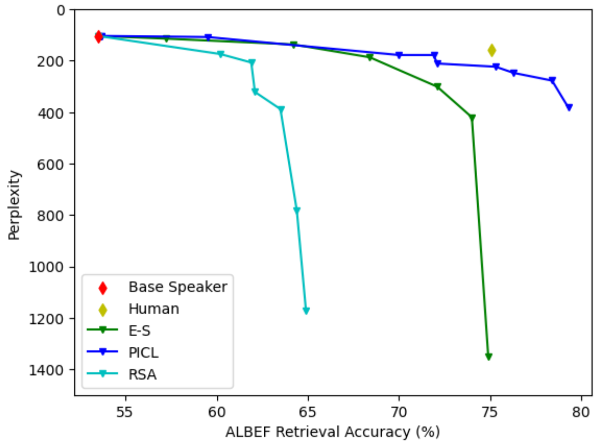


Figure 2: Informativeness (retrieval accuracy) versus fluency (perplexity) trade-off on ImageCoDe valid set. Compared with previous methods, our proposed PICL approach achieves a better trade-off between fluency and informativeness.

Video vs. static images set Table 2 also illustrates the large performance gap of ALBEF retrieval accuracy between video frames and static images, which is is consistent across all method-generated and human-written captions. This gap confirms that video frames pose much greater challenges to current vision-language models than static pictures for both retrieving target images and generating contextual and informative descriptions.

Ablation Studies 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 us-

	all	video	static
PICL	38.1	27.7	77.3
- incremental	32.0	23.2	65.4
- distractors	28.5	20.1	57.5

Table 4: Informativeness ablation experiment: results of the proposed approach on ImageCoDe test set. To investigate the role of incremental inference in PICL, we evaluate "- incremental" that only conducts CLIP scoring and reranking on full captions generated by the speaker model only. To quantify the effect of reasoning over the context in PICL, we experiment with "- distractor" in which only the target image is included during inference.

ing only the base speaker S_0 to generate a set of complete and context-agnostic captions, and using CLIP to score these entire captions. To allow CLIP to choose from the same number of candidates as each step in incremental inference, we perform beam search with S_0 using a beam width of N (subsection 3.3) to generate N 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 significantly on either of the two variations, suggesting that both the incremental inference and grounding to distractors are vital components for pragmatic reasoning in PICL.

6 Conclusion

In this paper, we study grounding language to visual context through the lens of pragmatics, with a focus on contrastive captioning. We propose an incremental pragmatic inference approach with a CLIP listener, which combines the strength of previous approaches that conduct incremental pragmatic reasoning with a separately modeled listener. Experimental results on a challenging dataset show that the proposed approach could generate captions that are highly informative without much loss of fluency. In the future, we plan to conduct human evaluations to verify the proposed method could generate high-quality captions that are discriminative for humans.

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.

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.

Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641.

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.

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

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.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint*, arXiv:2107.07651.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021b. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*.

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

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.

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.

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.

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*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.