Fine-grained Image Captioning with CLIP Reward

Jaemin Cho1  Seunghyun Yoon2  Ajinkya Kale3  Franck Dernoncourt2
Trung Bui2  Mohit Bansal1
1UNC Chapel Hill  2Adobe Research  3Adobe Inc.

{jincho, mbansal}@cs.unc.edu  {syoon, akale, franck.dernoncourt, bui}@adobe.com

Abstract

Modern image captioning models are usually trained with text similarity objectives. However, since reference captions in public datasets often describe the most salient common objects, models trained with text similarity objectives tend to ignore specific and detailed aspects of an image that distinguish it from others. Toward more descriptive and distinctive caption generation, we propose using CLIP, a multimodal encoder trained on huge image-text pairs from web, to calculate multimodal similarity and use it as a reward function. We also propose a simple finetuning strategy of the CLIP text encoder to improve grammar that does not require extra text annotation. This completely eliminates the need for reference captions during the reward computation. To comprehensively evaluate descriptive captions, we introduce FineCapEval, a new dataset for caption evaluation with fine-grained criteria: overall, background, object, relations. In our experiments on text-to-image retrieval and FineCapEval, the proposed CLIP-guided model generates more distinctive captions than the CIDEr-optimized model. We also show that our unsupervised grammar finetuning of the CLIP text encoder alleviates the degeneration problem of the naive CLIP reward. Lastly, we show human analysis where the annotators strongly prefer the CLIP reward to the CIDEr and MLE objectives according to various criteria.

1 Introduction

Describing an image with its detailed and distinguishing aspects is crucial for many applications, such as creating text keys for the image search engine and accessibility for the visually impaired. Standard deep learning approaches train an image-conditioned language model by maximizing the textual similarity between generated and reference captions (Vinyals et al., 2015; Xu et al., 2015; Rennie et al., 2017; Anderson et al., 2018). However, the reference captions of public datasets often describe only the most prominent objects in the images. This makes models trained to maximize textual similarity with reference captions tend to generate less distinctive captions that ignore the fine detailed aspects of an image that distinguishes it from others.

To alleviate the problem, we propose to use CLIP (Radford et al., 2021), a multimodal encoder model trained on large image-text data (mostly English) collected from the web, by using its similarity scores as rewards (Sec. 3.1). In addition, we propose a CLIP text encoder finetuning strategy with synthetic negative caption augmentation to improve the grammar of captioning model, without any extra text annotations (Sec. 3.2). Note that our approach completely eliminates the need for reference captions during reward computation. We illustrate our approach at Fig. 1. To comprehensively evaluate descriptive captions, we also introduce FineCapEval, a new dataset that measures captioning in diverse aspects: overall, background, object, and relation between objects (Sec. 4).

In our experiments on the MS COCO (Lin et al., 2014) dataset, we show that the captions of models trained with CLIP reward are more distinctive.
and contain more detailed information compared to the captions from CIDEr (Vedantam et al., 2015)-optimized models. CLIP-guided captions even achieve higher text-to-image retrieval performance than reference captions that are originally paired with images. We also show that our text encoder finetuning significantly improves caption grammars by removing degeneration artifacts such as word repetition. In fine-grained caption evaluation with FineCapEval and human analysis, we show that our CLIP-based rewards outperform text similarity objectives by a large margin in all categories.

2 Related Works

Image Captioning Metrics. Traditionally, captions have been evaluated with similarity metrics based on n-grams or scene graphs, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2015). However, such metrics often fail to capture paraphrased expressions due to the limited number of reference captions or scene-graphs. To address the problem, recent works including BERTScore (Zhang et al., 2019), ViLBERTScore (Lee et al., 2020a), UMIC (Lee et al., 2021), and CLIPScore (Hessel et al., 2021), propose using relevance scores computed by language or multimodal models pretrained on large data.

Objectives for Image Captioning. Standard deep learning-based image captioning approaches train models with a maximum likelihood estimation (MLE) objective. Ranzato et al. (2016) point that MLE suffers from an exposure bias problem.2 To address exposure bias, Bengio et al. (2015) propose a curriculum learning strategy called scheduled sampling. Ranzato et al. (2016) propose to train models by directly maximizing the text similarity between the generated and reference captions with REINFORCE (Williams, 1992). Rennie et al. (2017); Luo (2020) propose self-critical sequence training (SCST) approach by normalizing rewards to stabilize the high variance of rewards.

As illustrated in Fig. 2, de facto standard reward function for captioning is text similarity between generated and reference captions. Recent studies have found that reference-trained captioning models often neglect important information from images (Dai et al., 2017; Wang et al., 2017). Lee et al. (2020b) use accuracy of an visual question answering model as a reward, encouraging models to generate captions that include information sufficient to answer a visual question. Dai and Lin (2017); Luo et al. (2018); Liu et al. (2018) use image-text retrieval model’s self-retrieval score as a reward and combine them with metrics based on n-grams, encouraging captioning models to generate captions that are distinctive to each input image.

Note that these works require a careful balance between self-retrieval and text similarity objectives for stable training. In contrast, with the CLIP text encoder finetuning (Sec. 3.2), our approach eliminates the need for reference caption and text similarity metrics for the reward computation.

3 Methods

3.1 CLIP-guided Image Captioning

We propose using the CLIP (Radford et al., 2021) image-text similarity score to guide a image captioning model. Following Hessel et al. (2021), we use CLIP-S as our reward: 

\[
\text{CLIP-S}(I, c) = w \ast \max(f^I(I) f^T(c), 0),
\]

where \(I, c\) are the image and caption, \(f^I, f^T\) are the CLIP image and text encoders, and \(w = 2.5\). By learning to maximize the image-text similarity of the contrastive model, image captioning models are encouraged to generate captions that contain more distinctive information about the input image. Fig. 1 (a) illustrates this training strategy.

Following Rennie et al. (2017), we optimize our captioning model \(P_\theta(c|I)\) with REINFORCE (Williams, 1992) with a self-critical baseline. We approximate the gradient of expected reward for the generated caption \(\hat{c}\), where the reward of the beam search is normalized with the baseline reward \(b\) from greedy decoding \(\hat{c}_{\text{greedy}}\):

\[
\nabla_\theta \mathbb{E}_{\hat{c} \sim P_\theta(c|I)} [R(I, \hat{c})] \approx \nabla_\theta \mathbb{E}_{\hat{c} \sim P_\theta(c|I)} [R(I, \hat{c})]
\]

Figure 2: Comparison of different reward types for image captioning: (a) previous approaches with text similarity reward, such as CIDEr (Vedantam et al., 2015); (b) our image-text similarity reward based on CLIP.
We introduce FineCapEval, a new dataset for fine-grained caption evaluation in four different aspects. To construct FineCapEval, we collect 500 images from the MS COCO (Lin et al., 2014) test2015 split and Conceptual Caption (Sharma et al., 2018) val split, respectively. Then, for each image, we ask 5 human annotators to write phrases of 1) background, 2) objects (and their attributes; i.e., color, shape, etc.), 3) relation between objects (i.e., spatial relation), and 4) a detailed caption that includes all three aspects. See details of data collection process in appendix. In total, FineCapEval consists of 1,000 images with 5,000 annotations for each of the four criteria. In Table 1, we show samples from the FineCapEval dataset.

### 5 Experiments

We compare different reward configurations: MLE, CIDEr, CLIP-S, CIDER+CLIP-S, and CLIP-S+Grammar. Following previous work, we conduct experiments on the MS COCO (Lin et al., 2014) English captioning dataset with Karpathy split (Karpathy and Fei-Fei, 2015). We evaluate the model with n-gram based metrics, embedding based metrics, text-to-image retrieval scores, and Cider-Eval. We also perform a human evaluation with five criteria to understand the human preference for the generated captions in various aspects.

#### Model Architecture and Training

We use CLIP-Res50 transformer (Shen et al., 2022) as our captioning model architecture. The model consists of CLIP-Res50 for visual feature extraction and a transformer (Vaswani et al., 2017) encoder-decoder for conditional language model. We resize images in 224x224 to extract 2048-dimensional visual features. The transformer consists of 6-layer encoder and 6-layer decoder. We train our model with MLE
Table 2: Captioning performance of different rewards on MS COCO Karpathy test split. *The first caption out of 5 reference captions is used to calculate retrieval scores. R@K refers to the recall-K of the reference image. \( R_{word} \) refers to the word-level recall for background (Bg.), object (Obj.) and relation (Rel.) criteria (see Sec. 4 for details).

| Reward                  | N-gram based | Embed based | Text-to-Image Retrieval | FineCapEval |
|-------------------------|--------------|-------------|-------------------------|-------------|
|                         | BLEU-4 | CIDEr | METEOR | ROUGE-L | BERT-S | CLIP-S | RefCLIP-S | Overall | Bg | Obj | Rel |
| Reference captions      |         |         |         |         |         |         |           |         |    |    |    |
| MLE                     | 32.5    | 110.3   | 27.2    | 55.2    | 0.937   | 0.758   | 1.12      | 21.8    | 45.6 | 58.0 | 13.5 | 11.6 | 13.0 | 19.8 |
| CIDEr                   | 38.2    | 124.9   | 28.7    | 58.5    | 0.942   | 0.759   | 1.13      | 20.9    | 45.6 | 58.2 | 12.8 | 13.1 | 23.1 | 22.4 |
| CLIP-S                  | 6.2     | 11.2    | 18.7    | 31.6    | 0.882   | 0.870   | 1.17      | 42.5    | 71.6 | 82.2 | 13.9 | 20.8 | 26.4 | 24.9 |
| CIDEr+CLIP-S            | 37.7    | 124.6   | 28.8    | 58.3    | 0.941   | 0.772   | 1.14      | 24.4    | 30.2 | 63.1 | 13.0 | 13.0 | 23.4 | 21.7 |
| CLIP-S+Grammar          | 16.9    | 71.0    | 24.9    | 47.3    | 0.924   | 0.793   | 1.15      | 35.8    | 64.0 | 75.8 | 19.3 | 21.8 | 25.5 | 27.5 |

6 Results and Discussions
6.1 CLIP Guides Distinctive Captions
In Table 2, the models with CLIP-S and CLIP-S+Grammar rewards achieve higher image-text metrics (CLIP-S / RefCLIP-S) and text-to-image retrieval scores than baselines. Interestingly, their retrieval scores are even higher than the retrieval score with reference captions. This shows the distinctiveness of their generated captions. For image (a) in Table 3, our model with CLIP-S+Grammar reward describes the rainy weather with ‘wet’, while the model with CIDEr reward does not describe it. Our models with CLIP-S and CLIP-S+Grammar rewards score lower text similarity metrics (n-gram metrics and BERT-S) than the model with CIDEr reward. However, the low scores on these reference-based metrics can be addressed by that models with CLIP-S and CLIP-S+Grammar rewards often generate captions that include fine-grained information that is not even present in the reference captions. For image (b) in Table 3, CLIP-S+Grammar model describes ‘blue sign’ of the restaurant, whereas none of the reference captions mentions them.

6.2 Finetuning CLIP Text Encoder Improves Grammar
Table 3 shows that the degeneration (e.g. repetition of words) of the CLIP-S reward is successfully...
### Table 3: Captions generated by models with different rewards on MS COCO Karpathy test split images.

| Criteria | CLIP-S + Grammar | Win | Lose | Tie |
|----------|------------------|-----|------|-----|
| Overall  | v.s. MLE         | 49.0| 41.8 | 9.2 |
|          | v.s. CIDEr       | 51.0| 30.8 | 18.2|
| Background| v.s. MLE         | 52.8| 35.0 | 12.2|
|          | v.s. CIDEr       | 53.9| 25.4 | 20.6|
| Object   | v.s. MLE         | 52.0| 36.6 | 11.4|
|          | v.s. CIDEr       | 55.2| 32.8 | 12.0|
| Attribute| v.s. MLE         | 57.2| 36.8 | 6.0 |
|          | v.s. CIDEr       | 55.8| 37.2 | 7.0 |
| Relation | v.s. MLE         | 44.6| 44.2 | 11.2|
|          | v.s. CIDEr       | 49.2| 39.6 | 11.2|

Table 4: Human pairwise preference evaluation results.

mitigated by adding the grammar reward (CLIP-S+Grammar). Table 2 shows that adding grammar reward significantly increases all text similarity metrics (e.g., +60 for CIDEr).

### 6.3 Fine-grained Caption Evaluation

FineCapEval. The four right columns of Table 2 show that CLIP-S and CLIP-S+Grammar significantly outperform CIDEr on all four criteria of FineCapEval: overall, background, object, relation. The gap is smallest in the object criterion, which implies that MS COCO reference captions describe more object information than background or relation between objects.

Human Evaluation. Table 4 shows human evaluation results on five criteria: overall, background, object, attribute, relation. We sample 50 captions from model trained with CLIP-S+grammar reward (ours), CIDEr reward and MLE baseline using 50 images from Conceptual caption (Sharma et al., 2018) val split. For each of the five criteria, we ask 10 human annotators to select a better caption between ours and another method. On all criteria, human annotators strongly prefer captions with CLIP-S+Grammar rewards over CIDEr and MLE baseline.

### 7 Conclusion and Future Directions

We introduce a novel training strategy for image captioning models by maximizing multimodal similarity score of CLIP and finetuning its text encoder to improve grammar. The use of CLIP reward eliminates the need for reference captions and their bias for the reward computation. We also introduce FineCapEval, a dataset for fine-grained caption evaluation. We demonstrate the effectiveness of our proposed method based on improvements in text-to-image retrieval, FineCapEval, and human evaluation on fine-grained criteria along with qualitative examples. Future works involve finetuning CLIP reward models with desired writing styles for different applications and improving the synthetic augmentation process by using external data suitable for grammars with advanced linguistics expertise.

### 8 Ethical Considerations

The CLIP models that we used are trained on millions of web image-text pairs. Birhane et al. (2021) shows that such large-scale datasets often contain explicit and problematic image-text pairs. As the CLIP model card\(^6\) suggests, the use of CLIP reward to train image captioning models is intended as a research output, and any deployed use case of the models is out of scope.

Our captioning models and CLIP models are trained on English datasets; its use should be lim-

\(^6\)https://github.com/openai/CLIP/blob/main/model-card.md
itted to English language use cases. As our proposed method is not limited to English and is easily extended to other languages, future work will explore the extensions in various languages.

Acknowledgements

We thank the reviewers for their valuable comments. This work was partially done while JC was interning at Adobe Research and later extended at UNC, where it was supported by ARO Award W911NF2110220, DARPA MCS Grant N66001-19-2-4031, and NSF-CAREER Award 1846185. The views contained in this article are those of the authors and not of the funding agency.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In ECCV.

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 CVPR.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In ACL Workshop.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In NIPS, pages 1–9.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In ICCV.

Bo Dai and Dahua Lin. 2017. Contrastive Learning for Image Captioning. In NIPS.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-Free Evaluation Metric for Image Captioning. In EMNLP.

Andréj Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In CVPR.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC : An Unreferenced Metric for Image Captioning via Contrastive Learning. In ACL.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020a. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In EMNLP Workshop.

Kenton Lee, Ming-wei Chang Jonathan, and H Clark Regina. 2020b. CapWAP: Captioning with a Purpose. In EMNLP.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In ACL Workshop.

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 ECCV.

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 ECCV.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Ruotian Luo. 2020. A Better Variant of Self-Critical Sequence Training.

Ruotian Luo, Gregory Shakhnarovich, Scott Cohen, and Brian Price. 2018. Discriminability Objective for Training Descriptive Captions. In CVPR, pages 6964–6974.

Kishore Papineni, Salim Roukos, Todd Ward, and Wj Wei-jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In ACL.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chana, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In NIPS Workshop.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Jong Wook, Kim Chris, Hallacy Aditya, Ramesh Gabriel, Goh Sandhini, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In ICML.

Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In ICLR, pages 1–15.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical Sequence Training for Image Captioning. In CVPR.
Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhehui Yao, and Kurt Keutzer. 2022. How Much Can CLIP Benefit Vision-and-Language Tasks? In ICLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In NIPS.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In CVPR.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In CVPR.

Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. 2017. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In NIPS.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. In EMNLP.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In ICML.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In ICLR.

In this appendix, we include more example image captioning with different rewards (Sec. A), implementation details (Sec. B), FineCapEval details (Sec. C), human evaluation details (Sec. D), and the license for the datasets and models used in this project (Sec. E).

A More Image Captioning Examples

We provide more image captioning examples using different reward functions in Table 5. Overall, the captions from the model with CLIP-S+Grammar reward provide 1) more descriptive than the captions from the CIDEr model and reference captions, and 2) more grammatically correct than the captions from the model with CLIP-S reward.

B Implementation Details

Negative Caption Generation. In Alg. 1, we show Python implementation of the negative text generation (Sec. 3.2) for grammar finetuning. In summary, we generate negative captions using one of the operations: repeat, remove, insert, swap, shuffle on the original captions.

Evaluation Scripts. We use pycocoevalcap for MS COCO caption evaluation metrics such as CIDEr. We use BERTScore official repo with roberta-large model to calculate BERT-S. We report the evaluation script number from single run (single weight initialization), as we did not observe meaningful score fluctuation across multiple runs in our initial experiments.

C FineCapEval Details

Data Collection. To create a fine-grained description of the image, we ask annotators to write a caption that should describe target images’ 1) background, 2) objects and their attributes (i.e., color, shape, etc.), and 3) the relationship between the objects if any (i.e., spatial relation). Furthermore, we ask the annotators to write metadata containing which words/phrases in their writing belong to the three criteria. We also provide annotators with guidelines in writing a caption as follows: 1) There should be a single sentence describing the image. 2) The image may be a photo, an illustration or a pure background. 3) Pay close attention to local and global events in the image. 4) Descriptions should be at least ten words for each image. 5) Avoid the subject description of the image (i.e., a dog runs “very fast”, a man feels “successful”). 6) Avoid known entities such as specific locations (i.e. Eifel Tower), time (i.e., 4 pm), event (i.e., Halloween), proper name. 7) In describing people, use only man/woman/boy/girl if clear; otherwise, use person/child. All annotators are hired by a professional crowdsourcing platform TELUS. The crowdsourcing company obtained consents from the crowdworkers before the annotation process and conducted the ethical reviews. We collect English captions and all the annotators are native English speakers living in the US. We pay 5,400 USD, including 1) caption creation (5k samples) and 2)
| Image | Reward | Captions |
|-------|--------|----------|
| (a)   | CIDEr  | a group of boats parked in the water on a lake |
|       | CLIP-S | several rows of boats parked near a canal mountains horizon area and a mountain horizon area horizon ear motion |
|       | CLIP-S+Grammar | a lot of boats parked on the grass next to the lake with the hills behind |
|       | Reference Captions | A blue boat docked on a green lush shore. A small marina with boats docked there. A group of boats sitting together with no one around. Some boats parked in the water at a dock. Boats sitting around the side of a lake by a tree. |
| (b)   | CIDEr  | a zebra standing in the snow next to a brick wall |
|       | CLIP-S | a adult zebra wearing black and grey stripes standing near a brick wall area area with grey stance position stance |
|       | CLIP-S+Grammar | a large black and grey zebra standing together in the snowy ground next to a stone |
|       | Reference Captions | A zebra is standing outside in the snow. One zebra standing in snow near a stone wall. A zebra is standing in a snowy field. A zebra stands in snow in front of a wall. A zebra standing alone in the snow with a stone block wall and wooden fence behind it. |
| (c)   | CIDEr  | a black dog sitting next to a plate of food |
|       | CLIP-S | black black dog with macaroni and macaroni plate with pasta and pasta on a wooden floor plate position position position |
|       | CLIP-S+Grammar | a black dog sitting next to a plate of food on the wood floor |
|       | Reference Captions | Shaggy dog gets dinner served on a plate. A small black dog standing over a plate of food. A small dog eating a plate of broccoli. A black dog being given broccoli to eat. There is a dog staring at a plate of food. |
| (d)   | CIDEr  | two elephants standing next to a tree in a zoo |
|       | CLIP-S | two adult and baby elephant near a tree enclosure area with a tree area enclosure motion stance ear stance |
|       | CLIP-S+Grammar | a large elephant playing with a tree in the dirt field with rocks behind it |
|       | Reference Captions | An elephant standing under the shade of a tree. An elephant standing in the middle of a rocky environment. An elephant is alone in a wooded enclosure. An elephant standing in a shaded clearing in a wooded area. An elephant walks alone past some big rocks boulders in an open field. |
| (e)   | CIDEr  | a group of people riding bikes down a city street |
|       | CLIP-S | several cyclists moving and bicycles near a restaurant and a blue advertisement outside a red brick building motion stance p |
|       | CLIP-S+Grammar | a group of people riding their bikes on the busy street with a blue sign |
|       | Reference Captions | People on bicycles ride down a busy street. A group of people are riding bikes down the street in a bike lane. Bike riders passing Burger King in city street. A group of bicyclists are riding in the bike lane. Bicyclists on a city street, most not using the bike lane. |
| (f)   | CIDEr  | a man riding a bike next to a train |
|       | CLIP-S | older adult male riding a bicycle near a red and commuter train passing a train station motion stance ear stance |
|       | CLIP-S+Grammar | a person walking on a bike next to a red passenger train on the road |
|       | Reference Captions | A man on a bicycle riding next to a train. A person is riding a bicycle but there is a train in the background. A red and white train and a man riding a bicycle. A guy that is riding his bike next to a train. A man riding a bike past a train traveling along tracks. |
| (g)   | CIDEr  | a window of an airport with planes on the runway |
|       | CLIP-S | several rows of planes parked outside a terminal window area with fog outside a terminal window motion position area motion |
|       | CLIP-S+Grammar | a lot of airplanes parked on a wet airport terminal |
|       | Reference Captions | An airport filled with planes sitting on tarmac. The view of runway from behind the windows of airport. A truck driving towards some planes parked on the runway. Planes on a wet tarmac unloading at arrival gates. Window view from the inside of airplanes, baggage carrier and tarmac. |

Table 5: More captions generated by models with different rewards on MS COCO Karpathy test split images.

**Quality Assurance Process**

A quality assurance process that manually examines 50% of the created caption by different workers.

**Word-level Recall $R_{word}$**

In Alg. 2, we show Python implementation of word-level recall $R_{word}$. In summary, $R_{word}$ measures how many words from each of the reference phrases are included in a generated caption on average.

**Human Evaluation Details**

We conduct pairwise evaluation of human preference, as shown in the Sec. 5. For each image, we show two captions generated from two models: ours (CLIP-S + Grammar) and the baseline (MLE/CIDEr). A human worker selects a caption that better describes the image in terms of five criteria: overall, background, object, attribute, and relation. For each criterion, we use 50 images from
from random import randint, choice, shuffle

def repeat(tokens, n_max_gram=3, n_max_repeat=3):  # repeat n-grams
    n_gram = randint(1, n_max_gram)
    repeat_idx = randint(0, len(tokens) - n_gram)
    repeated = tokens[repeat_idx:repeat_idx+n_gram]
    n_repeat = randint(1, n_max_repeat)
    for _ in range(n_repeat):
        insert_idx = randint(0, len(tokens))
        tokens = tokens[:insert_idx]+repeated+tokens[insert_idx:]
    return tokens

def remove(tokens, n_max_gram=3):  # remove n-grams
    n_gram = randint(1, n_max_gram)
    remove_idx = randint(0, len(tokens) - n_gram)
    tokens = tokens[:remove_idx] + tokens[remove_idx+n_gram:]
    return tokens

def insert(tokens, vocab, n_max_tokens=3):  # insert tokens
    n_insert_token = randint(1, n_max_tokens)
    for _ in range(n_insert_token):
        insert_idx = randint(0, len(tokens) - 1)
        insert_tok = choice(vocab)
        tokens = tokens[:insert_idx] + [insert_tok] + tokens[insert_idx:]
    return tokens

def swap(tokens, vocab, n_max_tokens=3):  # swap tokens
    n_swap_tokens = randint(1, n_max_tokens)
    for _ in range(n_swap_tokens):
        swap_token_idx = randint(0, len(tokens) - 1)
        swap_token = choice(vocab)
        while swap_token == tokens[swap_token_idx]:
            swap_token = choice(vocab)
        tokens[swap_token_idx] = swap_token
    return tokens

def _shuffle(tokens):  # shuffle tokens
    shuffle(tokens)
    return tokens

def generate_negative_text(text, vocab):  # main function
    tokens = text.split()
    neg_type = choice(['repeat', 'remove', 'insert', 'swap', 'shuffle'])
    if neg_type == 'repeat': tokens = repeat(tokens)
    elif neg_type == 'remove': tokens = remove(tokens)
    elif neg_type == 'insert': tokens = insert(tokens, vocab)
    elif neg_type == 'swap': tokens = swap(tokens, vocab)
    elif neg_type == 'shuffle': tokens = _shuffle(tokens)
    return " ".join(tokens)
Algorithm 2 Python implementation of word-level recall $R_{\text{word}}$ computation (main paper Sec. 5)

```python
def calculate_word_recall(pred_id2sent, gt_id2phrases):
    ""
    pred_id2sent: dict of generated captions (dict[int, str])
    gt_id2phrases: dict of reference phrases (dict[int, list[str]])
    """
    n_total = 0
    total_score = 0
    for id, gt_phrases in gt_id2phrases.items():
        pred_sent = pred_id2sent[id]
        score = 0
        for gt_phrase in gt_phrases:
            word_score = 0
            for gt_word in gt_phrase.split():
                if gt_word in pred_sent:
                    word_score += 1
            score += word_score / len(gt_phrase.split())
        score /= len(gt_phrases)
        total_score += score
        n_total += 1
    word_recall = total_score / n_total * 100
    return word_recall
```

Figure 3: The screenshot of human evaluation process for ‘object’ criterion (main paper Sec. 5).
FineCapEval, and the two options are randomly and evenly shuffled. We also provide ‘Tie’ option to choose when the two captions are equally good or bad. For each criterion, we recruit 10 annotators 1) who are located in the Great Britain or the United States 2) HIT approval rate above 80% and 3) Number of HITs approved greater than 1000, from Amazon Mechanical Turk. We pay the annotators 0.03 USD per selection, which roughly corresponds to 11 USD/hour. In Fig. 3, we provide the screenshot for ‘object’ criterion for example.

E Licenses
For all artifacts, we remain within their respective license agreements. Here, we list the licenses:

- **MS COCO - CC 4.0** - https://cocodataset.org/#termsofuse
- **Conceptual Captions** - https://github.com/google-research-datasets/conceptual-captions/blob/master/LICENSE
- **CLIP - MIT** - https://github.com/openai/CLIP/blob/main/LICENSE
- **CLIP-ViL - MIT** - https://github.com/clip-vil/CLIP-ViL/blob/master/LICENSE