Improving Image Captioning by Mimicking Human Reformulation Feedback at Inference-time

Anonymous ACL submission

Abstract

Incorporating automatically predicted human feedback into the process of training generative models has attracted substantial recent interest, while *feedback at inference time* has received less attention. The typical feedback at training time, i.e., preferences of choice given two samples, does not naturally transfer to the inference phase. We introduce a novel type of feedback – caption reformulations – and train models to mimic reformulation feedback based on human annotations. Our method does not require training the image captioning model itself, thereby demanding substantially less computational effort. We experiment with two types of reformulation feedback: first, we collect a dataset of human reformulations that correct errors in the generated captions. We find that incorporating reformulation models trained on this data into the inference phase of existing image captioning models results in improved captions, especially when the original captions are of low quality. We apply our method to non-English image captioning, a domain where robust models are less prevalent, and gain substantial improvement. Second, we apply reformulations to style transfer. Quantitative evaluations reveal state-of-the-art performance on German image captioning and English style transfer, while human validation with a detailed comparative framework exposes the specific axes of improvement.[1]

1 Introduction

There is a growing interest in feedback models that approximate human feedback *during training* of generative models. While resulting generative models achieve improved performance on automatic metrics and human evaluations (Ouyang et al., 2022; Faltings et al., 2023), the use of feedback models during training requires the generative model to be trained or at least fine-tuned.

[1]Our code and data are available here: github.com/uriberger/re_cap.git

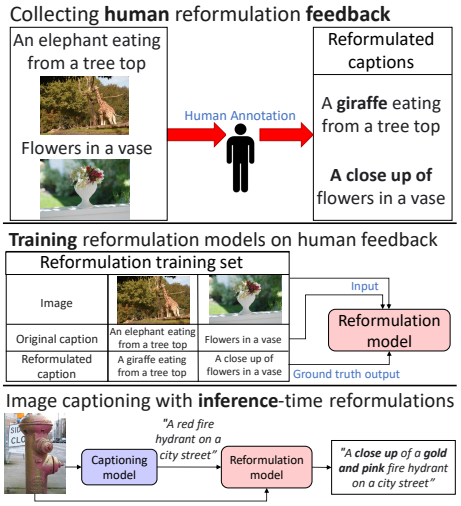


Figure 1: Our proposed method with reformulation for improved factuality as an example. Top: Collecting human-written reformulations of model captions. Center: Using the collected data to train models to generate reformulations, given an input image and original caption. Bottom: Combining an off-the-shelf captioning model (no training) with our reformulation model, to adapt generated captions at inference time.

The use of such feedback models *during inference* poses no such requirement, but was nevertheless generally overlooked by previous studies. One reason for this limited interest is the type of feedback that existing feedback models predict: comparative feedback, e.g., by predicting human preference for one of two generated candidate outputs (as used in Reinforcement Learning from Human Feedback, Stiennon et al., 2020). While this type of feedback naturally translates into a reward function to be used during training, it is less clear how to employ it at inference time when model parameters are fixed.

We bridge this gap by proposing a novel type of feedback, namely **reformulation** (see Figure 1 and Section 2). We focus on the image captioning task, since it provides a good testing ground for adapt-

ing a general model to fit specific user intent. For example, one user may require the captions to describe the colors in the image while another would focus on a specific style of generated captions.

When providing reformulation feedback for model-generated captions, human annotators receive an image and a model-generated textual description as input, and subsequently produce text that is as similar as possible to the input text but also incorporates an additional desired attribute, e.g., improved factuality or a desired style (Fig. 1, top). We train models to mimic this type of feedback (Fig. 1, center) and integrate them into the inference phase of off-the-shelf image captioning models (Fig. 1, bottom).

A small amount of data (a few thousand samples, as demonstrated in our experiments in Sections 3 and 4) is sufficient to train a reformulation model that once trained, can be applied to any captioning model without further training it, making reformulation models a much more efficient alternative to training-time feedback models that require to retrain the captioning model.

To study the benefits of this type of feedback, we focus on two reformulation attributes. First, we train models to rewrite the input caption with improved factuality (Section 3). We collect English reformulation data by asking human annotators to correct errors in generated captions while making minimal changes, and use this data to train a reformulation model. We then use the reformulation model on captions generated by off-the-shelf English models. We show that the automatic reformulation process notably improves captions generated by weaker models, while careful analysis including a fine-grained human evaluation paradigm reveals that, similar to human reformulations, the most notable factor in the improvement of the automatic reformulation process is adding missing information. To further investigate the utility of our method in domains where existing models are weak ("challenge domains"[2]), we propose a cross-lingual pipeline for reformulation in German image captioning and show notable improvement, achieving state-of-the-art performance in German image captioning on the Multi30k (Elliott et al., 2016) dataset.

Second, we cast caption style transfer as a reformulation task (Section 4). We use existing parallel stylized and non-stylized caption data, and train a reformulation model to preserve the structure of the input caption while adapting its style to a given target. We build on powerful but style-agnostic captioning models using style-reformulation at inference time, achieving state-of-the-art performance on the FlickrStyle dataset on automatic metrics, while our human evaluation paradigm confirms that the reformulated captions are more stylized than competitive baselines.

Each of the reformulation attributes studied in this work (improved factuality and style transfer) reveal a use case for our method. In the first, the goal of the user is to improve captioning models in challenge domains. This is accomplished by selecting a reformulation attribute that will improve the quality of the captions (factuality in our case) and apply corresponding reformulation models to a weak captioning model. In the second case the user aspires to generate high-quality captions in a specific style, and therefore utilizes a robust captioning model to generate high quality captions and then change their style using reformulation.

2 Modeling Reformulation Feedback

In this Section, we define our notion of reformulation feedback. A human annotator observes an image and a caption describing the image, and produces a caption that 1) incorporates some desired attribute (e.g., factuality or some desired style) , and 2) is as similar to the input caption as possible. Since these two criteria are in conflict we emphasize that the first requirement is obligatory, but annotators should make minimal changes to achieve it. Note that reformulation may be applied in any generation task, but here we focus only on image captioning.

In this study we focus on two attributes of reformulation feedback: improved factuality (Section 3) and style transfer (Section 4).

Reformulation model. Recent research examined frameworks of multimodal input (image+text) and unimodal output (text), demonstrating that fine-tuning a checkpoint that was pre-trained on general Vision-and-Language tasks is an effective approach (e.g., in Visual Qustion Answering, Chen et al., 2022). We follow this strategy by fine-tuning the pre-trained mPLUG (Li et al., 2022a) check-

[2]We define "challenge domains" as the (many) domains dominated by weaker models, e.g., low-resource scenarios or niche domains less amenable to established model architectures.

point[3] on reformulation data[4].

3 Reformulation for Improved Factuality

Captioning models in challenge domains, e.g., non-English captioning, tend to generate captions of lower quality compared to English captioning models. We propose to use reformulations to improve the factuality of models in these domains. In this Section, we study this use case. We first describe data collection and then apply our model to English and German image captioning.

3.1 Data Collection

Data. To generate an initial set of image captions, we use three publicly available captioning models, that vary in architecture, size and amounts of training data: BLIP (Li et al., 2022b), mPLUG (Li et al., 2022a), and ClipCap (Mokady et al., 2021). We randomly sample 1405 images from the test sets of MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014), and generate a caption with each model.

Annotation. Human annotators were shown an image and a model-generated caption, and asked to reformulate the caption so that (a) it is as similar as possible to the original caption and (b) any errors in the original caption are corrected (if any errors were present).

Annotators were instructed to consider a wide range of errors in their feedback, including hallucinations (describing elements that are not present in the image), partial descriptions (failing to describe a key element in the image) and replacements (using an incorrect word to describe an element in the image).

We use Amazon Mechanical Turk to recruit annotators. For the full details on annotator recruitment, guidelines and payment, see Appendix B.

Data analysis. In 864 samples (16.6%) the annotators chose not to change the original caption. The mean Levenshtein distance[5] is 4.79. Additionally, we sample 100 random captions that were changed by the annotators and classify the changes to the element changed (object, action, object attribute, setting, other) and the nature of the change (add, replace, remove, rewrite). 'Setting' changes are changes in the setting of the caption (e.g., adding the location in which the caption takes place is classified as 'add setting'). 'Other' captures any change that is not covered by the first four elements. If most of the objects, actions and attributes in the reformulated caption differ from those of the original caption, we classify the change as 'rewrite'. Results in Table 1 show that *object* is the most frequently changed element: in 51% of the captions an object was added, replaced or removed. The most common type of change (applied in 73% of the captions) is adding information. We find that all the annotators' modifications were valid[6].

	Add	Replace	Remove	Rewrite	Total
Object	24	24	3	–	51
Action	11	7	0	–	18
Attribute	12	0	3	–	15
Setting	26	3	0	–	29
Other	0	9	0	–	9
Total	73	43	6	15	

Table 1: Statistics for reformulations of 100 random labeled data points. One reformulation may contain several operations.

3.2 Improved Factuality for English Image Captioning

In this section we experiment on English data. We use off-the-shelf captioning models on well known captioning datasets and reformulate the generated captions using the model described in Section 2 trained on the data described in Section 3.1. To test the reformulation model on data both from a familiar and an unfamiliar distribution, we use the models (BLIP, ClipCap, mPLUG) and datasets (MSCOCO, Flickr30k) used to generate the reformulation training data (Section 3.1) excluding the images that were already presented to the reformulation model during training, as well as models (GIT: Wang et al. 2022, vit_gpt2: Kumar 2022) and datasets (XM3600: Thapliyal et al., 2022) with which the reformulation model is unfamiliar.

As described above, in this use case we expect to improve the factuality of weaker models. We therefore mainly focus on relatively weak captioning models: we use the pretrained only (not finetuned) checkpoint of mPLUG, the base version of GIT, and ClipCap and vit_gpt2 which are realtively old and small models. For completion we also use one strong model, the finetuned checkpoint of BLIP.

[3] We also experimented with BLIP, but mPLUG performed significantly better.

[4] For the full training details see Appendix A.

[5] Minimum number of words needed to be added, removed or replaced to get from original to reformulated caption.

[6] For the list of manually examined captions, see supplementary materials.

3.2.1 Automatic Evaluation

We present the change in performance for different metrics in Table 2. We use the commonly used (e.g., Li et al., 2022a,b) metrics BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). In addition to these 4 general metrics we report the performance on different types of sentential elements, provided by the SPICE metric (objects, relations, attributes, size words, color words, cardinality words). This allows us to observe a change in performance specifically regarding the sentential elements that our reformulation models are trained to address (Table 1).

For the weaker models (mPLUG, ClipCap, GIT, vit_gtp2) we see an improvement across all datasets and general metrics. For BLIP we see a minor decrease in performance on BLEU-4 and CIDEr, and a minor increase on METEOR and SPICE. Turning to SPICE components, improvement was observed across all weaker models, datasets and components, except for Color with GIT on Flickr30k. For BLIP, improvement was observed in most configurations, but most notably in color words. Therefore, our reformulation model is particularly well-suited for domains characterized by a lack of robust models, such as non-English image captioning, on which we focus in Section 3.3.

Figure 2 shows examples where SPICE scores were notably higher after reformulation, for each SPICE element. In accordance with Table 1, most of the improvement originates from information that was added during reformulation.

3.2.2 Human Evaluation

To qualitatively evaluate the changes during reformulation, we propose a fine-grained human evaluation paradigm. For each of the models we randomly sample images from each of the datasets (17 from MSCOCO and Flickr30k, 16 from XM3600 to a total of 50 per model), and present human annotators with the images along with the original caption and the reformulated caption, in randomly shuffled order and without indicating the source of each caption. As we expect to observe notable improvement for weaker models, we exclude BLIP from this analysis[7]. Three on-site annotators with high English proficiency assessed the captions. For each sample, the annotators answer the following questions, each related to one axis of caption quality (in bold):

- **Faithfulness:** Which caption includes less content that is not in the image?
- **Completeness:** Which caption covers more elements of the image being described?
- **Accuracy:** Which caption uses fewer incorrect words to describe one of the object/activities in the image?
- **Detail:** Which caption includes more *properties* (such as color or shape) of the main objects in the image?
- **Overall:** Which caption is the better description of the image?

For each question, the annotators were given three options (first caption is better, second is better, both are equal). If at least two annotators prefer one of the captions along an axis, we mark the caption as 'better'. Otherwise both are considered 'equal'.

Figure 3 presents the results. Across all axes, reformulated captions are significantly (Sign test, $p < 0.05$) better than the original captions. Specifically, we see notable improvement in the overall quality (reformulated captions were better in 76%) and the completeness (46%) of the caption. This result is in line with the analysis presented in Table 1, where the most common feedback type was 'addition' of information to the original caption. The inter-annotator agreement using Fleiss' Kappa was $0.68, 0.68$ for completeness, overall (substantial agreement, Landis and Koch, 1977), and $0.56, 0.55, 0.53$ for faithfulness, detail, accuracy (moderate agreement).

3.3 Improved Factuality for Cross-Lingual Image Captioning

The last section demonstrated strong gains of our approach for weak off-the-shelf models. Acknowledging that image captioning models sharply drop in performance in languages other than English, we next investigate the use of English reformulation in a cross-lingual setup. We combine a German image captioning model with our reformulation model by generating German captions; translating the captions to English; reformulating them with our model; and translating back to German.

Data. We use Multi30k (Elliott et al., 2016), a large, non translated, German image caption dataset, which contains 30K/1K images for train/test, each with 5 captions. All images are taken from the Flickr30k dataset and all captions are generated by German native speakers.

[7] See Appendix E for a similar analysis for BLIP.

Dataset	Model	General metrics				SPICE components					
		B@4	M	C	S	Obj	Rel	Att	Car	Siz	Col
MSCOCO	ClipCap	6.3	3.3	7.2	6.5	7.6	7.6	15.5	194.6	1.8	19.8
	mPLUG	24.5	23.3	32.1	29.9	26.2	36.5	51.5	221.1	90.6	44.8
	GIT	81.3	41.3	57.4	42.4	34.0	93.4	70.9	58.1	53.3	31.6
	vit_gpt2	1.2	5.6	3.2	9.7	6.1	15.1	31.6	59.3	18.2	129.9
	BLIP	-5.2	0.6	-3.2	1.8	1.9	1.1	7.2	5.7	6.7	16.8
Flickr30k	ClipCap	21.3	10.3	30.9	16.6	13.7	20.4	37.6	115.7	10.3	30.8
	mPLUG	30.9	28.6	55.5	33.9	31.1	24.5	79.0	294.9	213.2	50.0
	GIT	45.2	32.5	19.2	27.6	24.0	155.4	19.0	45.7	87.9	-12.3
	vit_gpt2	20.8	17.1	34.9	25.9	19.0	117.1	76.1	121.4	35.0	160.3
	BLIP	-6.5	1.3	-2.2	1.6	1.5	2.2	5.5	8.3	-3.6	14.5
XM3600	ClipCap	14.6	8.4	21.3	14.0	12.6	15.1	27.0	122.0	0.0	17.8
	mPLUG	148.6	49.8	60.3	40.7	36.8	53.1	79.6	∞	143.9	62.0
	GIT	46.5	23.9	23.6	11.9	11.4	4.7	13.3	70.8	33.7	1.3
	vit_gpt2	32.3	18.3	34.1	21.8	17.6	77.3	67.8	89.3	72.7	89.1
	BLIP	-3.5	1.9	1.8	0.7	0.9	-1.6	3.7	-7.6	4.6	6.0

Table 2: Performance change after reformulation compared to raw model output on common metrics, datasets and models (in % of the recorded performance before reformulation). We observe major improvements in weaker models (ClipCap, mPLUG, GIT, vit_gpt2). Darker green (red) indicates higher improvement (deterioration). M: METEOR, C: CIDEr, S: SPICE. ∞ marks a configuration where the metric value before reformulation was 0.

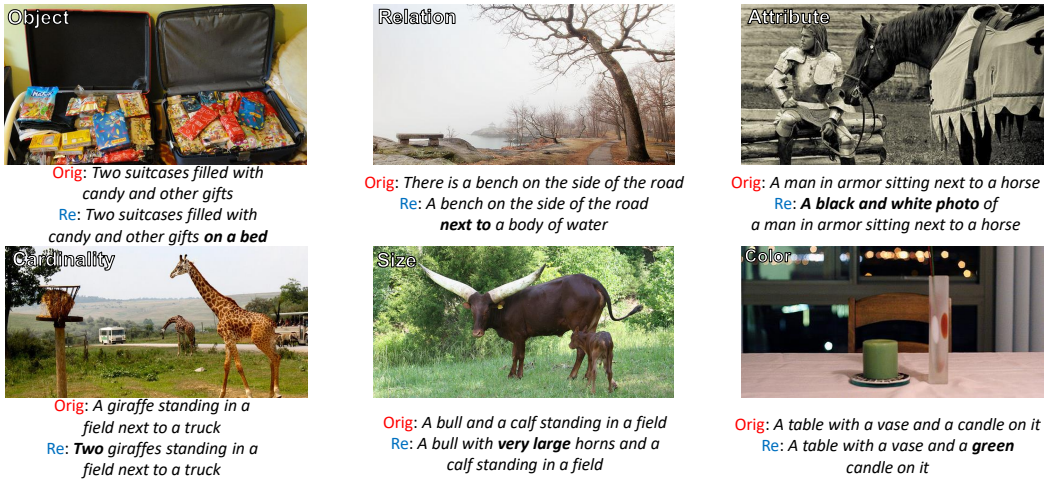


Figure 2: Examples in which the reformulated captions achieved better results than the original ones, in all SPICE elements. Orig: the caption generated by the model. Re: the reformulated caption.

Model. Due to a lack of a strong and publicly available pretrained image captioning model for German, we train our own model. We use the ClipCap model as it separates the text decoder from the image encoder, allowing us to straighforwardly incorporate a German decoder. We use the original ClipCap implementation[8] and change the text decoder to a German version of GPT2.[9] We refer to this model as **base**. We reformulate the captions generated by **base**, and refer to these as **base+re**. Following recent captioning works (Thapliyal et al., 2022; Ramos et al., 2023b), we use Google Translation API for all translations.

Baselines. First, to directly measure the performance gain of the reformulation pipeline, we use **base** as a baseline. Second, the mPLUG checkpoint on which the reformulation model is based (see Section 2) is in itself quite a capable captioning model. Consequently, given an input image and caption the reformulation model might ignore the input caption and generate its own caption. To make sure this is not the case, we also generate English captions using the reformulation model by providing an image and an empty caption as input, and translate these captions to German (**tran**). Finally, we present results reported by recent German image captioning studies: Dual Attention (**DA**, Jaffe, 2017), Cycle Consistency (**CC**, Wu et al., 2019) and Multi-Objective Optimization (**MOO**, Wu et al., 2022). We report the same metrics as in Section 3.2.1 except SPICE which, to the best of our knowledge, is not available for German.

[8] github.com/rmokady/CLIP_prefix_caption
[9] huggingface.co/dbmdz/german-gpt2

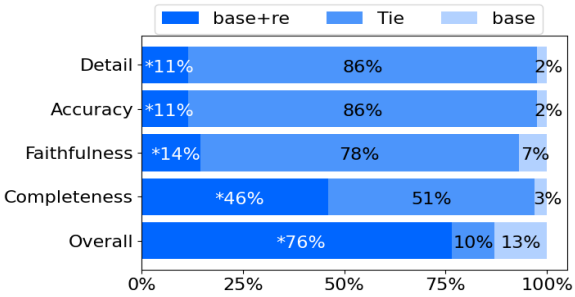


Figure 3: Results for human evaluation on different axes. We show proportions of preferences for generated captions without (base) and with (base+re) reformulations, and ties. * indicates a significant difference between base and base+re (Sign test; $p < 0.05$).

	B@4	METEOR	CIDEr
base	12.8 ± 0.3	18.6 ± 0.2	39.2 ± 1.6
tran	14.3	**20.3**	46.0
DA	16.0	17.8	30.8
CC	15.9	17.8	31.0
MOO	16.5	17.9	33.8
base+re	**16.8 ± 0.1**	20.1 ± 0.1	**51.4 ± 0.6**

Table 3: German Results on Multi30k test set. Results for the models that we train (base, base+re) are averaged over 3 random initializations and we report the standard deviation. For each metric, the best result is bolded.

Results. Results in Table 3 show that **base+re** outperforms all other methods in BLEU-4 and CIDEr, while **tran** achieve the best result in METEOR, though by a small margin. The improvement over **base** emphasizes the power of the reformulation pipeline, while the improvement over **tran** suggests that providing the reformulation model with a reasonable caption is an important factor in the success of the reformulation process. We also note the improvement over previous state-of-the-art studies. We partially attribute this to the use of the strong German GPT2 model (since **base** outperforms previous models on two metrics), but reformulation contributes notable value, as evidenced by the superiority of **base+re** over **base**.

3.3.1 Human Evaluation

To better understand the improvement reported by the automatic metrics, we follow the same protocol as in Section 3.2.2. The annotation was conducted by two on-site German native speakers with an inter-annotator agreement score (measured by Cohen's Kappa) of at least 0.54 across all axes.

Results are presented in Figure 4. We notice that while in English improvement was most significant

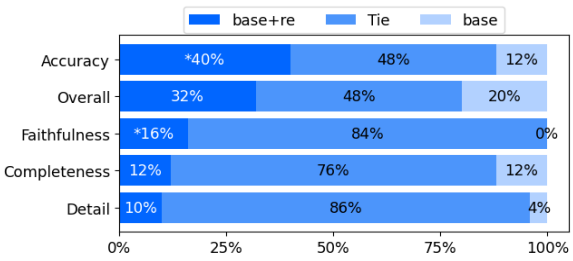


Figure 4: Results for human evaluation on different axes of German generated captions. We show proportions of preferences for generated captions without (base) and with (base+re) reformulations, and ties. * indicates significance as in Figure 3.

in terms of Completeness (Figure 3), in German the most significant axes are Faithfulness and Accuracy. We hypothesize that the captions produced by the German base model contain many errors and the focus of the reformulation process is therefore on fixing the errors, while errors in the English generated captions are rare and thus the focus is on adding new information. We corroborate this hypothesis by computing the mean caption length before and after reformulation for English (44.6 → 49.3) and German (57.3 → 54.1). See Appendix D for examples.

4 Reformulation for Style Transfer

We study the generalizability of reformulation feedback modeling by focusing on a second reformulation attribute: the style of the caption, i.e., the reformulation should adapt the style while making minimal changes.

4.1 Dataset

We use the FlickrStyle (Gan et al., 2017) dataset. FlickrStyle contains humorous and romantic captions for 7000 images from Flickr30K. Importantly, the annotators were instructed to generate the captions based on existing captions from Flickr30K. We follow Wang et al. (2023) and randomly split the data to 6000 train images and 1000 test images.

4.2 Method

We train a reformulation model for a given style as follows. First, for each caption in FlickrStyle we identify the original caption in Flickr30K on which that caption is based by measuring the string overlap of the stylized caption with each of the original captions of the same image, and selecting the caption with the largest overlap. Next, we fine-tune a reformulation model as described in Section 2, with

Style	Method	B@1	B@3	M	C
Humorous	CapDec	29.4	8.8	13.2	55.1
	SAN	29.5	9.9	12.5	47.2
	TridentCap	30.6	11.2	12.8	56.6
	BLIP	29.6	11.0	14.4	**73.9**
	BLIP+re	**33.7**	**11.7**	**14.8**	72.0
Romantic	CapDec	27.9	8.9	12.6	52.2
	SAN	30.9	10.9	13.0	53.3
	TridentCap	31.9	11.4	13.4	60.4
	BLIP	28.5	11.2	14.3	72.0
	BLIP+re	**35.1**	**13.0**	**15.4**	**74.6**

Table 4: Results for stylized image captioning on FlickrStyle. B@n: BLEU-n, M: METEOR, C: CIDEr. For each style and metric, the best result is in bold.

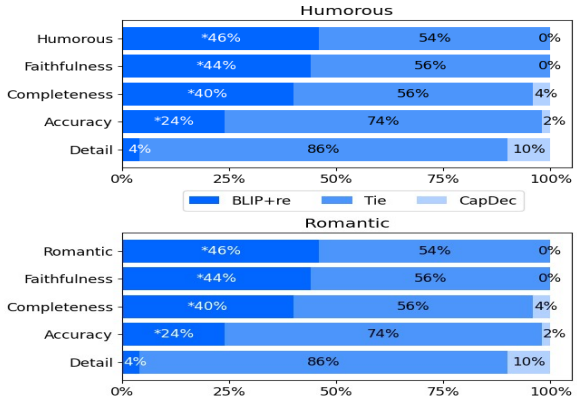


Figure 5: Results for human evaluation on different axes of stylized captions. We show proportions of preferences for the baseline (CapDec) and BLIP reformulated (BLIP+re) captions, and ties. * indicates significance as in Figure 3.

Figure 6: Examples of original captions and reformulated captions for humor and romantic reformulation.

the original caption as the input and the stylized caption as the ground-truth output.

4.3 Models

We use BLIP as the captioning model (**BLIP**) and for each style, we reformulate the BLIP captions using a reformulation model trained to transfer captions to the style in question (**BLIP+re**). Note that vanilla BLIP does not generate stylized captions (i.e., is expected to perform poorly on this task). As baselines, we present results from previous studies: CapDec (Nukrai et al., 2022), SAN (Li et al., 2021), and TridentCap (Wang et al., 2023).

4.4 Automatic evaluation

Results are presented in Table 4. We follow the convention from previous stylized image captioning studies and report Bleu-1, Bleu-3, METEOR and CIDEr. Our method achieves state-of-the-art results for both styles, and we attribute this improvement to the strong captions generated by the BLIP model (in the humor style **BLIP** even outperforms **BLIP+re** in the CIDEr metric). This unveils an issue in automatic evaluation: vanilla BLIP outperformed the baselines though it clearly does not generate stylized captions (see Figure 6 for examples). The same may be true for **BLIP+re**. Thus, we conduct human evaluation to ensure that captions generated by **BLIP+re** are indeed stylized.

4.5 Human Evaluation

We again use our human evaluation scheme (Section 3.2.2) to compare to previous baselines. We compare to CapDec[10], since we found no available codebases for TridentCap and SAN. We ask the first 4 questions from Section 3.2.2 (Faithfulness, Completeness, Accuracy, Detail) and add a style-related question: *Which caption is more {humorous,romantic}?*

Results are presented in Figure 5. Our method improves over the baseline not only in the quality of captions, but also in generating stylized captions, significantly in both styles. Annotators agreement (Fleiss' Kappa) values were $\kappa = 0.59, 0.51, 0.48, 0.34$ for Faithfulness, Style, Completeness, Accuracy (the axes where reformulated captions were better), and $\kappa = 0.44$ for Detail.

5 Related Work

We classify related work by the type of feedback (human/model-generated) and the phase in which the feedback is applied (training/inference): human feedback during training (Section 5.1), model-generated feedback during training (Section 5.2), and model-generated feedback during inference (Section 5.3, our study is included in this category).

[10] github.com/DavidHuji/CapDec

5.1 Human Feedback during Training

A large volume of previous studies collect human feedback and use it directly to improve training. Most studies focus on either comparisons (which of two candidate texts is better) or ratings as a reward signal in reinforcement learning for various tasks, e.g., dialogue (Jaques et al., 2020), machine translation (Kreutzer et al., 2018a) or semantic parsing (Lawrence and Riezler, 2018). Kreutzer et al. (2020) fine-tune a generative model with on-line feedback from human annotators.

Other types of training-time feedback include natural language comments (Campos and Shern, 2022). Most similar to our reformulation feedback are edits (Liu et al., 2022; Lu et al., 2023), where humans change an incorrect response generated by dialog models into a correct response. While these studies use feedback directly we train models to predict it, avoiding the necessity to collect annotations each time the feedback is used.

Image captioning. Several previous studies trained captioning models with raw human feedback. Shen et al. (2019) propose a model that generates a caption and subsequently generates a question pertaining to factual information within that caption. This question is then answered by a human. No questions are generated during inference. Seo et al. (2020) use human ratings of captions as rewards in a reinforcement learning framework.

Ling and Fidler (2017) are most similar to our study: they compare training captioning models on human-generated captions with training on reformulations, showing that the latter improves standard metrics. However, they use human-generated reformulations during training while we use model-generated reformulations during inference.

5.2 Model-Generated Feedback during Training

Several works train feedback models, but use these models during training, again predominantly focusing on comparisons or ratings feedback. Early studies (Christiano et al., 2017; Ibarz et al., 2018) use feedback models to train agents in simulated environments and games. Others use feedback models to train language models for specific tasks such as summarization (Ziegler et al., 2019; Stiennon et al., 2020), machine translation (Kreutzer et al., 2018b) and visual storytelling (Hsu et al., 2021). Recently, feedback models were used in training of general-purpose large language models (e.g. GPT-4, OpenAI, 2023). Most relatedly, Faltings et al. (2023) investigate reformulation feedback models, but only during training. Finally, constitutional AI (Bai et al., 2022) use similar ideas to train non-harmful models but use model feedback rather than human feedback.

5.3 Model-Generated Feedback during Inference

Most similar to ours, some previous studies apply feedback models at inference time. Hsu et al. (2019) train models to predict human post-edits of model generated text but focus on the visual storytelling task. Ramos et al. (2023a) apply metrics trained to predict human rating feedback for reranking model outputs in machine translation. To the best of our knowledge, we are the first to use feedback models at inference time for image captioning.

6 Discussion

Despite the recent success of incorporating (models of) human feedback as a training signal, using feedback during inference has received little attention. We presented a novel approach – reformulation feedback at inference time – and applied it to the task of image captioning.

We refrain from comparing our approach to a baseline of fine-tuning the captioning model directly on the corrected captions for two reasons. First, even if such baseline would induce better results, our method's advantage is efficiency, as reformulation models are trained once and can be combined with any base model architecture, while fine-tuning would be performed on any new model. Second, this baseline is not applicable in our cross-lingual use-case (Section 3.3), as the corrected captions are in English.

We've studied two use-cases for our method: improving captioning models in challenge domains (Section 3.3) and generating high quality stylized captions (Section 4). Both can be extended in future work: captioning models in other challenge domains (e.g., medical image captioning) can gain improved factuality, while robust models can be utilized to generate captions in other styles (e.g., sentimental captions). Taken together, our work contributes to the active areas of learning from human feedback, and efficient adaptation of powerful LLMs to diverse tasks.

Limitations

Data collection. While our method requires less computational resources compared to previous studies (since only the feedback model is trained rather than the generative model), it requires more human resources for annotation. Simpler types of feedback (e.g., the common comparative feedback) require less effort and time per sample than reformulation, while some studies (e.g. Ramos et al., 2023a) refrain from explicitly collecting any feedback data, by using publicly available human annotations that were originally collected for a different purpose (e.g., to train evaluation metrics).

Cross-lingual reformulation. The pipeline suggested in Section 3.3 for cross-lingual reformulation (generation of captions in the target language, translation into English, reformulation, translation back into the target language) depends on the existence of a decent base captioning model in the target language and good translation models from/to English. If the base captioning model in the target language generates poor captions, the reformulated captions will be no better than captions generated in English and translated to the target language (i.e. the **tran** baseline discussed in Section 3.3). If there are no strong translation models from/to English, the quality of captions would decrease in every translation step in the pipeline, resulting in poor captions. Future work may address training non-English reformulation model to bridge the second gap.

Variation in annotation conditions. Previous studies (Khashabi et al., 2022) show that human annotations may vary drastically when basic conditions change, e.g., on different days or even at a different time during the day. Since reformulation models are trained on such annotations, this may have a significant impact on the model. We did not take this into account in our data collection and usage.

Ethics Statement

In our data collection in Section 3.1 we collect no identifying data on the annotators. For existing datasets, we use publicly available resources in accordance with their license agreements. The datasets are fully anonymized and do not contain personal information about the caption annotators or any information that could reveal the identity of the photographed subjects.

As with other methods for modifying model outputs, our approach can be used to transfer toxic text to non-toxic text, or vice versa. Additionally, the reformulation data that was collected and presented in Section 3 may contain social biases. Along with the publication of our model and data, we will include a model card (Mitchell et al., 2019) which reports standard information regarding the collected data, training methods and intended use.

This work was approved by the <Removed for anonymization> Committee for the Use of Human Subjects in Research in <Removed for anonymization>.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jon Ander Campos and Jun Shern. 2022. Training language models with language feedback. In *ACL Workshop on Learning with Natural Language Supervision. 2022.*

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *ArXiv preprint*, abs/2209.06794.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the*

5th Workshop on Vision and Language, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Felix Faltings, Michel Galley, Baolin Peng, Kianté Brantley, Weixin Cai, Yizhe Zhang, Jianfeng Gao, and Bill Dolan. 2023. Interactive text generation. *ArXiv preprint*, abs/2303.00908.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 955–964. IEEE Computer Society.

Chi-yang Hsu, Yun-Wei Chu, Ting-Hao Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4443–4453, Online. Association for Computational Linguistics.

Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao Huang. 2019. Visual story post-editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6581–6586, Florence, Italy. Association for Computational Linguistics.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8022–8034.

Alan Jaffe. 2017. Generating image descriptions using multilingual data. In *Proceedings of the Second Conference on Machine Translation*, pages 458–464, Copenhagen, Denmark. Association for Computational Linguistics.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3985–4003, Online. Association for Computational Linguistics.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. 2022. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11444–11458, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144, Lisboa, Portugal. European Association for Machine Translation.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018a. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018b. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics.

Ankur Kumar. 2022. The illustrated image captioning using transformers. *ankur3107.github.io*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. biometrics, 159-174.

Carolin Lawrence and Stefan Riezler. 2018. Improving a neural semantic parser by counterfactual learning from human bandit feedback. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1820–1830, Melbourne, Australia. Association for Computational Linguistics.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. 2022a. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *ArXiv preprint*, abs/2205.12005.

Guodun Li, Yuchen Zhai, Zehao Lin, and Yin Zhang. 2021. Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5363–5372.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Huan Ling and Sanja Fidler. 2017. Teaching machines to describe images with natural language feedback.

In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5068–5078.

Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony Liu, and Soroush Vosoughi. 2022. Second thoughts are best: Learning to re-align with human values from text edits. *Advances in Neural Information Processing Systems*, 35:181–196.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2023. Towards boosting the open-domain chatbot with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078, Toronto, Canada. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *ArXiv preprint*, abs/2111.09734.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. *ArXiv preprint*, abs/2211.00575.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André FT Martins. 2023a. Aligning neural machine translation models: Human feedback in training and inference. *ArXiv preprint*, abs/2311.09132.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023b. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. *ArXiv preprint*, abs/2305.19821.

Paul Hongsuck Seo, Piyush Sharma, Tomer Levinboim, Bohyung Han, and Radu Soricut. 2020. Reinforcing an image caption generator using off-line human feedback. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2693–2700. AAAI Press.

Tingke Shen, Amlan Kar, and Sanja Fidler. 2019. Learning to caption images through a lifetime by asking questions. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 10392–10401. IEEE.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *ArXiv preprint*, abs/2205.12522.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Lanxiao Wang, Heqian Qiu, Benliu Qiu, Fanman Meng, Qingbo Wu, and Hongliang Li. 2023. Tridentcap: Image-fact-style trident semantic framework for stylized image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yike Wu, Shiwan Zhao, Jia Chen, Ying Zhang, Xiaojie Yuan, and Zhong Su. 2019. Improving captioning for low-resource languages by cycle consistency. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 362–367. IEEE.

Yike Wu, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, and Zhong Su. 2022. When pairs meet triplets: Improving low-resource captioning via multi-objective optimization. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–20.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv preprint*, abs/1909.08593.

A Model Training Details

A.1 Reformulation Models

We now specify the details of the reformulation models trained in Sections 3 and 4.2.

We use the VQA training pipeline from the official mPLUG code base.[11] We use the default hyperparameters, and fine-tune the mplug.en.base checkpoint for 8 epochs with the AdamW optimizer and learning rate of 3e-5. Models were trained on an Nvidia RTX a5000 GPU and each training session took less than an hour. Models contain 350M parameters.

A.2 German Captioning Model

We train the model discussed in Section 3.3 for 10 epochs with the AdamW optimizer and learning rate of 2e-5. The model was trained on an Nvidia RTX a5000 GPU and training took 4 hours to complete. The model contains 156M parameters.

B Data Collection

In this section we thoroughly discuss the data collection process briefly discussed in Section 3.1.

We use Amazon Mechanical Turk to recruit annotators. As a first filter we require native-speaker level proficiency in English. Next, we publish a qualification task and filter the annotators. Finally, after each batch of annotation, we sample 20 annotated samples to ensure the quality of annotations, and inform annotators if a wrong annotation has been made.

Annotators were paid 0.1$US per annotation. Early experiments indicated that a single reformulation annotation takes 5 to 30 seconds. The expected hourly wage exceeds the US minimum wage which ranges between 8$US and 15$US.

We provide the following annotation guidelines:

- In this task, you will be presented with images together with a textual image description.
- Your task is to reformulate the description so that (a) it is as similar as possible to the original (b) all errors from the original descriptions are fixed (if any errors exist).
- If the original description is too bad to fix, please write a completely new description.

Subsequently, annotators were shown several examples of reformulations.

C Used Packages

We used the following packages in our implementation:

- COCO-caption evaluation[12]: used for all evaluation metrics.
- statsmodel: used for sign-test[13] and Fleiss' Kappa[14] in the human evaluation sections.
- sklearn: used for Cohen's Kappa[15] in Section 3.3.

D More Examples

Figure 7 presents samples where the German captioning **base** model discussed in Section 3.3 generates caption with errors, which are fixed by the reformulation process.

E Analysis of BLIP Reformulation

We use the evaluation framework described in Section 3.2.2 on the BLIP model. We randomly sample 50 images from each of the MSCOCO and Flickr30k test sets for the evaluation.

Figure 8 presents the results. Across datasets, reformulated captions are more complete and detailed but less faithful and accurate. This result is in line with the analysis presented in Table 1, where the most common feedback type was 'addition' of information to the original caption. The reduction in accuracy and faithfulness shows that in some cases the added information was incorrect. However, annotators scored the reformulated captions as overall better in both datasets.

We find that reformulated captions are significantly (Sign test, $p < 0.05$) more detailed in MSCOCO, less faithful in Flickr30k, more complete in both datasets and overall better in both datasets ($p < 0.05$). We also compute inter-annotator agreement using Fleiss' Kappa: $\kappa =$

[11] github.com/alibaba/AliceMind

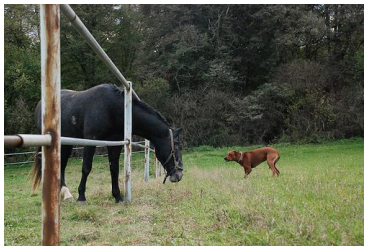
[12] github.com/tylin/coco-caption

[13] www.statsmodels.org/stable/generated/statsmodels.stats.descriptivestats.sign_test.html

[14] https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

[15] scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

base: *Ein Mann sitzt an einem Tisch vor einem Glas **Bier***
*(A man sits at a table in front of a glass of **beer**)*
base+re: *Ein Mann sitzt an einem Tisch vor einem Glas **Wein***
*(A man sitting at a table in front of a glass of **wine**)*



base: *Ein Hund steht an einem Zaun vor einem **Stall***
*(A dog stands at a fence in front of a **stable**)*
base+re: *Ein Hund steht an einem Zaun vor einem **Pferd***
*(A dog stands at a fence in front of a **horse**)*

base: *Ein Junge in roter Jacke und Helm sitzt auf einem **roten** Motorrad*
*(A boy in a red jacket and helmet sits on a **red** motorcycle)*
base+re: *Ein Junge in roter Jacke und Helm sitzt auf einem Motorrad*
(A boy in a red jacket and helmet sits on a motorcycle)

base: ***Zwei** Radfahrer fahren auf einer Brücke über einen Fluss*
*(**Two** cyclists ride on a bridge over a river)*
base+re: ***Drei** Radfahrer fahren auf einer Brücke über einen Fluss*
*(**Three** cyclists ride on a bridge over a river)*

Figure 7: Examples in which the reformulated captions fix errors in captions generated by the base model, for German image captioning. base: the caption generated by the base model. base+re: the reformulated caption.

$0.55, 0.47, 0.44$ for completeness, overall, detail (axes on which reformulated captions were better), and $\kappa = 0.37, 0.34$ for faithfulness, accuracy.

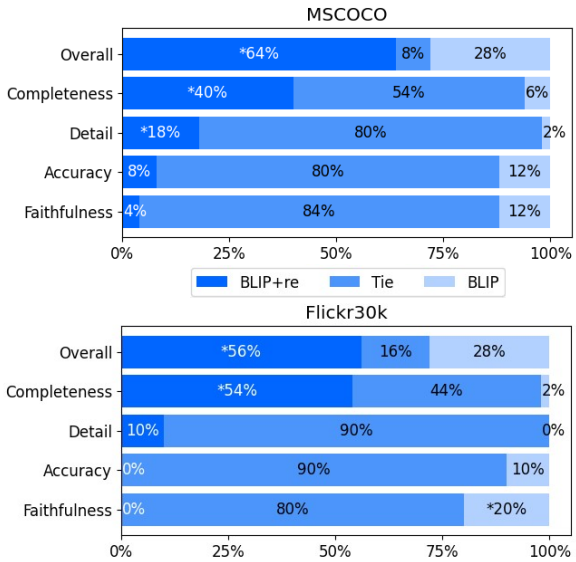


Figure 8: Results for human evaluation on BLIP in different axes. We show proportions of preferences for generated captions without (base) and with (base+re) reformulations, and ties. * indicates a significant difference between base and base+re (Sign test; $p < 0.05$).