# The Role of Data Curation in Image Captioning

**Wenyan Li**[1]**, Jonas F. Lotz**[1,2]**, Chen Qiu**[3]**, Desmond Elliott**[1]
[1]Department of Computer Science, University of Copenhagen
[2]ROCKWOOL Foundation Research Unit
[3]School of Computer Science and Technology, Wuhan University of Science and Technology
`{weli,jonasf.lotz,de}@di.ku.dk, chen@wust.edu.cn`

## Abstract

Image captioning models are typically trained by treating all samples equally, neglecting to account for mismatched or otherwise difficult data points. In contrast, recent work has shown the effectiveness of training models by scheduling the data using curriculum learning strategies. This paper contributes to this direction by actively curating difficult samples in datasets *without* increasing the total number of samples. We explore the effect of using three data curation methods within the training process: complete removal of a sample, caption replacement, or image replacement via a text-to-image generation model. Experiments on the Flickr30K and COCO datasets with the BLIP and BEiT-3 models demonstrate that these curation methods do indeed yield improved image captioning models, underscoring their efficacy.

## 1 Introduction

Image captioning is the task of generating grammatically correct and accurate descriptions of visual data, which involves understanding the identity of salient objects and their relationships (Bernardi et al., 2016; Baltrušaitis et al., 2018). While existing models have made significant progress on this problem, there remains an inherent challenge: how to address the variations in learning difficulty that arise from diverse image-caption pairs (Sharma et al., 2018; Schuhmann et al., 2021).

Image captioning models are usually trained by treating the entire training dataset equally, which overlooks the variations in the complexity of each data point. One attempt at addressing this issue has been to apply data filtering as a preprocessing stage to large-scale datasets to remove noisy data from the pretraining process (Li et al., 2022a; Nguyen et al., 2023). Several other image captioning techniques have relied on curriculum learning strategies (Bengio et al., 2009), which schedule the training data with increased levels of complexity, effectively adapting the learning process to the difficulty of the task (Liu et al., 2021; Dong et al., 2021; Zhang et al., 2022; Alsharid et al., 2021; Ayyubi et al., 2023). In this paper, we aim to answer a fundamental question: can image captioning models be improved by not only recognizing variations in the data but also actively curating difficult samples?

We introduce three data curation methods, each with the aim of improving the learning process while preserving the overall size of the training dataset. These methods include the complete removal of a sample, the replacement of captions, or the substitution of images using a text-to-image generation model. The targets of these methods are image-caption training samples that have unusually high losses with respect to the rest of the training dataset under the current model parameters. In other words, our approach focuses on the samples that are proving *difficult* to model (Bengio et al., 2009; Kumar et al., 2010).

The main findings of this paper are:

- Dynamic data curation enhances image captioning performance. The best strategy varies between datasets but is generalizable to different vision-language models.[1]

- The extent of curation is a critical factor and dataset dependent. We find that curating more than 50% of data negatively impacts the effectiveness of data curation.

- Image generation-based curation has potential benefits with specific techniques, but its potential benefit is limited by generation errors identified through a human study, which are not apparent from automatic evaluation metrics, such as CLIPScore (Hessel et al., 2021).

---

[1]We release the code for our curation framework at `https://github.com/lyan62/data-curation/`

## 2 Related work

**Data Curation in NLP** While still under-explored for image captioning, Rogers (2021) highlighted the importance of data curation for deep learning and NLP. Several studies have adopted data curation for large language models: Chen et al. (2023) developed a general text curation framework based on large language models; Kandpal et al. (2022) and Lee et al. (2022) discussed the impact of deduplication for training; Chang and Jia (2023) shows that careful curation alone can stabilize in-context learning.

**Image Captioning and Learning Strategies** Curriculum learning (Bengio et al., 2009) and self-paced learning (Kumar et al., 2010) are techniques that adjust the learning process based on variations in the learning samples, leveraging loss values to estimate model competence. For image captioning, several studies have introduced diverse learning techniques aimed at customizing the model training process in terms of sample difficulty, incorporating both textual and visual features (Alsharid et al., 2021; Dong et al., 2021; Zhang et al., 2022). Whereas these methods adjust model training using sorted data, our approach proposes an innovative perspective: adjusting training by curating data samples that exhibit outlier losses, while preserving the overall dataset size.

**Text-to-image Generative Models** Text-to-image generative models, including diffusion models (Song et al., 2021; Nichol and Dhariwal, 2021), have rapidly gained popularity and proven powerful. Although recent large-scale latent diffusion models excel in generating high-resolution images with artistic and photo-realistic qualities (Rombach et al., 2022; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), their application in multimodal tasks remains unexplored. Concurrently to our work, Azizi et al. (2023) and Jain et al. (2023) show that image classifiers can be improved by learning from augmented images generated by finetuned generative models; Xiao et al. (2023) and Caffagni et al. (2023) used generative models to augment the datasets used to train captioning models.

To the best of our knowledge, we are the first to explore how dynamic data curation approaches can impact downstream image captioning *without* scaling up existing datasets, and how text-to-image generative models can be applied in the process.

## 3 Data Curation for Captioning

Our main goal is to assess whether actively curating image-caption pairs during training can improve image captioning models. There are many reasons for the existence of difficult samples, including mismatches between the image-caption or inconsistencies between the image and caption (Atliha and Šešok, 2020), e.g. the caption includes mentions of entities that cannot be seen in the image. For clarity in what follows, let $\mathcal{D}$ be an image captioning training dataset with $K$ images, and let $\mathrm{I}_k$ be the $k$-th image. Each image is paired with $J$ captions; let $\mathrm{C}_k^j$ be $j$th caption of image $k$, and thus, let ($\mathrm{I}_k$, $\mathrm{C}_k^j$) be an image–caption sample.

### 3.1 Identifying the difficult samples

Inspired by scheduling in curriculum learning (Bengio et al., 2009; Kumar et al., 2010), we assume that difficult training samples can be automatically identified throughout the training process. We propose to use the captioning model $\mathcal{M}$ that is being trained on dataset $\mathcal{D}$ to automatically identify such samples. We can readily use this model to calculate the loss of each sample in $\mathcal{D}$ at any point in time, such as at the end of each epoch $t$: $\mathcal{L}_{\mathcal{M}}^t(\mathrm{I}_k, \mathrm{C}_k^j) \, \forall j, k$. The samples can be be ranked by their respective losses, providing candidates for samples that may benefit from data curation. In particular, the highest loss samples are targets for our data curation methods. We focus on samples with losses that are either two standard deviations from the mean, or the top $X\%$ highest loss samples. The data curation performs dynamic updates to the training dataset $\mathcal{D} \to \mathcal{D}_1 \to \cdots \to \mathcal{D}_T$. In this way, the training dataset is dynamically updated at the end of each epoch according to the model's current captioning capability at time $t$. We empirically observe that without data curation, the high-loss samples remain high-loss during five epochs of training.[2]

### 3.2 Curation approaches

We investigate three approaches to dynamically curate the high-loss image-caption pairs: REMOVAL, REPLACECAP, and REPLACEIMG. Figure 1 shows an overview of these approaches.

**REMOVE** The simplest approach to data curation is to remove the high-loss samples, preventing the samples from confusing the model. In REMOVE,

---

[2]The leftmost plot in Figure 5 shows the empirical distribution of losses in the training samples of the Flickr30K.
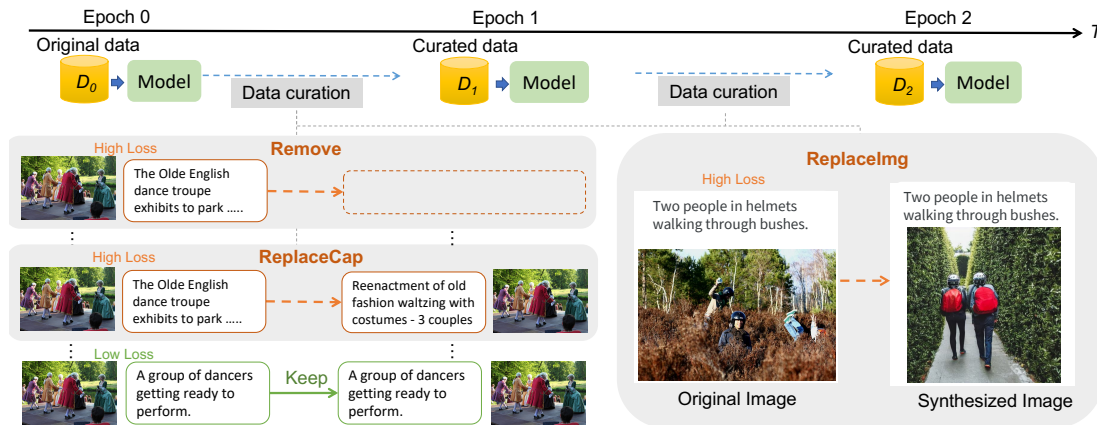
Figure 1: Overview of our data curation methods. For REMOVE, high loss image-text pairs are removed; for REPLACECAP, the image is paired with an alternative caption from the original dataset; for REPLACEIMG, captions of original images are used as prompts for text-to-image generation to synthesize new image–text pairs. We experiment with both options of replacing the image only, or pair another relevant caption to the synthesized image.

the high-loss samples are completely removed from the remainder of the training process, reducing the total number of image–caption training samples.

**REPLACECAP**  In REPLACECAP, we simply replace the caption in the image–caption sample with a different caption from the original dataset that describes the image, effectively creating a duplicate. With this method, the total number of samples used to train the model remains the same, as well as the total number of the unique images. This creates a control condition for our experiments. As an alternative, we also experiment with replacing the original caption with one generated by a language model, which we discussed in Section 6.

**REPLACEIMG**  In REPLACEIMG, we perform data curation using a text-to-image generative model. This has the benefit of training the model on the same total number of samples while exposing it to more unique images. In a rapid model-in-the-loop step, we use a text-to-image generation model to synthesize images based on the other sentences that describe the image. We integrate this into training as follows: Given an image $I_k$ in the training data and its captions $\{(I_k, C_k^1), \ldots, (I_k, C_k^J)\}$, we synthesize a new image $\hat{I}_k$ without increasing the total number of samples in the original dataset. Specifically, for image $I_k$, we replace an original high-loss sample $(I_k, C_k^j)$ with the synthesized image-text pair $(\hat{I}_k, C_k^j)$.

Given a set of captions that describe an image, there are several options for how to prompt the image generation model (Figure 11 in Appendix). We experiment with three options:

- Single caption: Each caption is used in isolation to generate a new image.

- Sentence-BERT selection: There is a lot of variety in how different captions describe the same image. Instead of using all captions, we can use a representative caption from the set. This is achieved using the Sentence-BERT (Reimers and Gurevych, 2019) model to find the caption that is closest to the average embedding of all captions.

- Concatenation: All five captions are concatenated as the text prompt for generation.

For all three approaches mentioned above, we can append an additional string to the prompt as a *styler* to force a specific style in the generated image (+Styler). The styler used here is: "national geographic, high quality photography, Canon EOS R3, Flickr".[3]  Some representative examples of images generated using this technique can be seen in Figure 13 in the Appendix.

## 4 Experimental Setup

### 4.1 Data & Metrics

We evaluate our data curation methods during fine-tuning on the widely used MS COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) datasets. We report results using the metrics of **B**LEU (Papineni et al., 2002), **M**ETEOR (Denkowski and Lavie, 2014), **C**IDEr (Vedantam et al., 2015), and **CLIPS**core (Hessel et al., 2021).

---

[3]The styler was chosen by inspecting the generated images, with a preference for photographic outputs and against "artistic" outputs, such as sketches and computer art.

| | Method | Ratio | BLIP | | | | BEiT-3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | B | M | C | CS | B | M | C | CS |
| Flickr30K | Baseline | - | 37.6 | 27.2 | 92.8 | 78.6 | 28.9 | 27.2 | 79.3 | **80.4** |
| | +Remove | 2 std | 38.6 | **27.4** | **95.8** | **79.2** | 31.4 | 27.1 | **83.7** | 80.0 |
| | +ReplaceCap | 1% | 37.9 | **27.4** | 94.5 | 78.9 | 29.6 | **27.5** | 80.1 | 80.3 |
| | +ReplaceImg | 40% | **39.0** | 27.3 | 95.7 | 79.1 | **32.0** | 26.9 | 82.4 | 79.1 |
| COCO | Baseline | - | 39.9 | 30.8 | 132.0 | 77.3 | 39.4 | 31.1 | 133.7 | 77.4 |
| | +Remove | 1% | 40.1 | 30.9 | 132.5 | 77.3 | 39.3 | 31.1 | 133.2 | 77.3 |
| | +ReplaceCap | 1% | **40.2** | 30.9 | 132.7 | 77.3 | 39.4 | 31.0 | 133.6 | 76.5 |
| | +ReplaceImg | 10% | **40.2** | **31.0** | **133.1** | 77.3 | **39.6** | **31.1** | **134.4** | **77.5** |

Table 1: Results of finetuning with our data curation methods compared to standard finetuning of BLIP and BEiT-3 on the Flickr30K and COCO datasets. We report **B**LEU, **M**eteor, **C**IDEr, and **CLIPS**core. Best scores are in **bold**.

## 4.2 Models & Implementation

**Image Captioning Models** We study the effectiveness of data curation with two state-of-the-art pretrained vision-language models – BLIP (Li et al., 2022a) and BEiT-3 (Wang et al., 2023).

We note that BLIP has a captioning and filtering (CapFilt) data augmentation process during its pretraining, where both components were finetuned on the COCO dataset. Therefore we use pretrained checkpoint $BLIP_{CapFilt}$ for Flickr30k and $BLIP_{base}$ for COCO in experiment, removing the effects of the CapFilt process. We finetune BLIP using a total batch size of 128 for 5 epochs on $4 \times$ A100 GPUs. The BEiT-3 base model is finetuned with the default setups: a total batch size of 256 for 10 epochs on $8 \times$ A100 GPUs.

**Curation Ratio** We tune the amount of data to be curated for each method on the validation data of each dataset using the BLIP model. See Section 6 for more discussion on the trade-off between the amount of data curation and model performance.

**REPLACEIMG Text-to-image Generation** For text-to-image generation in REPLACEIMG, we use the open source Stable Diffusion model (Rombach et al., 2022), which can generate images given a textual prompt. We finetune a Stable Diffusion v1.5 model, using the MS COCO (Lin et al., 2014) dataset with a prompt consisting of a concatenation of all 5 captions, for 15,000 steps with a constant learning rate of $1e{-}5$ and a batch size of 32. We experiment different versions of the released Stable Diffusion models and various techniques for
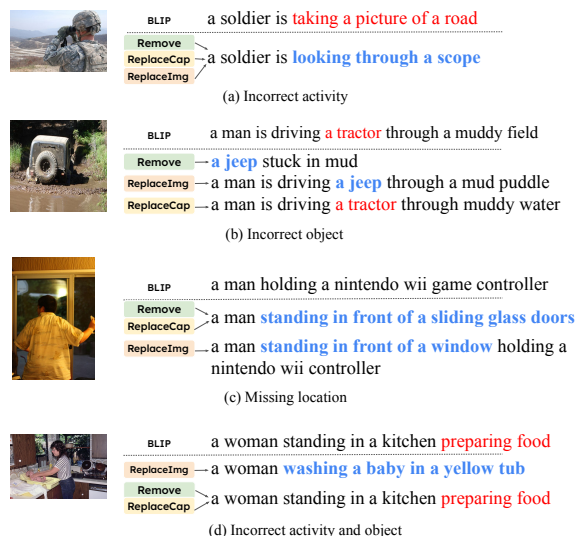


Figure 2: Qualitative examples from the COCO dataset of captions generated by the BLIP model (top), and the same models trained using our data curation methods (bottom). After curation, many of the errors (in red) can be avoided or fixed (in blue).

generating high-quality images for replacement.[4] We find that using a finetuned text-to-image model enhances image captioning performance. See Section 7 for further analysis and ablation.

## 5 Results

**Data curation improves captioning** Table 1 shows the results for the Flickr30K and COCO datasets with the BLIP and BEiT-3 models. The main conclusion is that better model performance

---

[4]It is also possible to use API-based models but we chose Stable Diffusion because (i) Stable Diffusion can be integrated directly into our training pipeline using the open source code. And (ii) we estimate that it would cost \$4,176 to run a single experiment on the Flickr30K dataset using DALL·E-2 as of Feburary 1st, 2024.
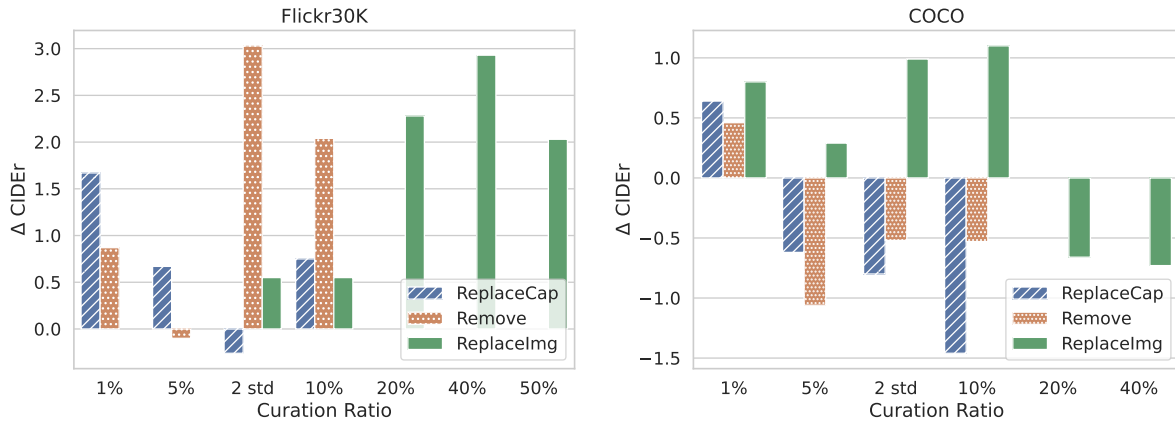
Figure 3: Effects of varying the amount of data curated. We observe that Flickr30K needs more curation (40% REPLACEIMG or 2 std REMOVE) than COCO (10% REPLACEIMG or 1% REPLACECAP). Flickr30K benefits more from removing high-loss training samples, indicating the original dataset may be noisier than MS COCO. For the 2 std approach, the number of samples curated is not fixed after each epoch and varies between 5% to 10%.

can almost always be achieved using data curation.

For Flickr30K, it can be seen that REMOVE (2 std) and REPLACEIMG (40%) perform similarly well with a 2.9–3 CIDEr points improvement. The REPLACECAP method only improves performance by 1.7 CIDEr points when applied to the top 1% of high-loss samples. For COCO, the best performing approach is REPLACEIMG with a curation ratio of 10%, bringing a 1.1 CIDEr point improvement over the baseline. REPLACECAP and REMOVE both work best when curating the top 1% of high-loss samples, bringing smaller improvements of 0.5–0.7 CIDEr points. Qualitative examples of the improvements can be seen in Figure 2.

**Generalization to different VL models** We also verify that our data curation methods generalize to other models by implementing them in the BEiT-3 model. More specifically, we used exactly the same curation ratio that gained improvements for BLIP. As shown in Table 1, where REMOVE is also the most efficient approach for better captioning on Flickr30K, and REPLACEIMG improves the most for COCO. This shows that the curation methods can be readily applied to other state-of-the-art vision-language models and the curation ratios are transferable. We note that since BEiT-3 includes COCO in pretraining, the REMOVE and REPLACECAP methods are not beneficial.

## 6 Discussion

**Curation amount matters** The amount of data curated is an important hyperparameter. In addi-
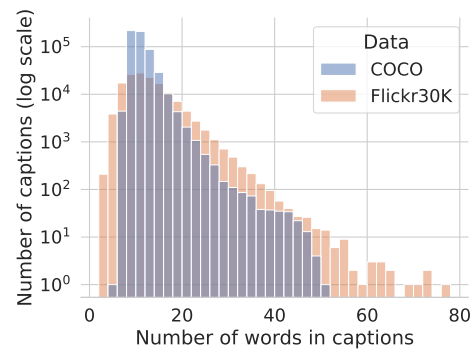


Figure 4: Distribution of caption lengths.

tion to the best results reported above, we present finer-grained results of varying the amount of data curation. For REMOVE and REPLACECAP, we explore curating the top 1%, 5% and 10% of high-loss samples. For REPLACEIMG, we explore 10%–80% curation ratios. In addition to fixed X% ratios, we also intereven on samples that have losses two standard deviations worse than the mean.

The results of this analysis are shown in Figure 3. While the effective curation ratio for different curation approach ranges from 1%-50% for Flickr30K, COCO benefits from REPLACEIMG on less than 10% of the top loss samples, and the effective curation ratio for REMOVE and REPLACECAP stops at 1%. This indicates that Flickr30K may contain more noisy samples than the MS COCO dataset. Compared to MS COCO, Flickr30K contains more samples with long captions (Figure 4), which may include overly-specific details that are inconsistent with other captions and are hard for the model to learn (Figure 12). Through our curation-based finetuning, these samples can be effectively iden-
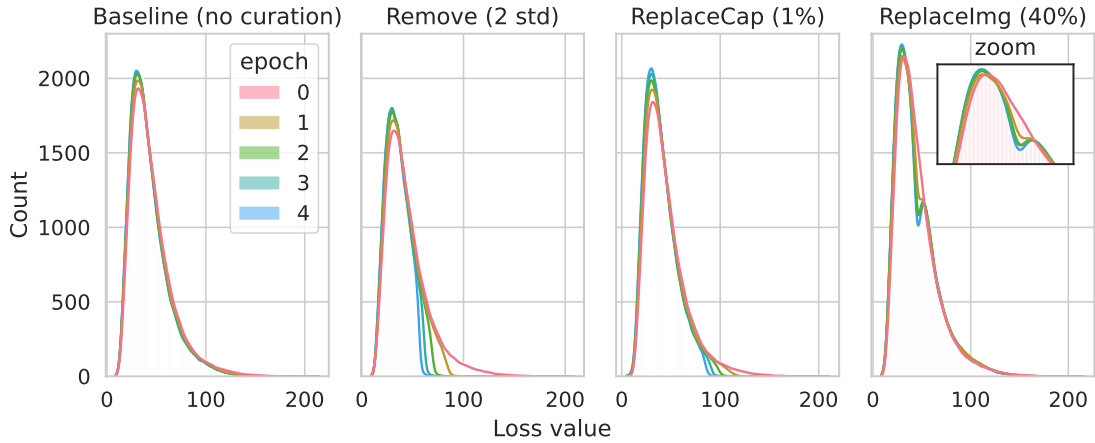
Figure 5: Different curation methods change the loss distribution of training samples over epochs for Flickr30K. In contrast, in the absence of data curation (the leftmost plot), high-loss samples consistently retain their high-loss status throughout the training process.
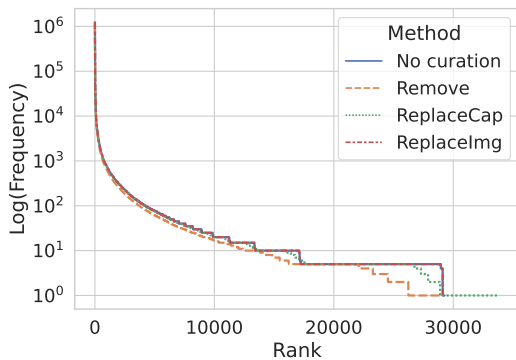


Figure 6: Zipfian distribution of words in Flickr30K training samples for different curation approaches. Note the clear changes made to the tail by REMOVE.



Figure 7: Dynamic versus static replacement for RE-PLACEIMG using BLIP on the Flickr30K dataset, as a function of the number of samples replaced.

tified, removed or replaced, which indicates that our method is efficient when training with noisy datasets. We note that curating more than 50% of the data does not benefit training and actually harms performance.

**Curation changes training distributions** We examine the loss distributions of training samples across epochs for each curation method to understand their impact on the training process (Figure 5). These losses are computed after each epoch using the current model parameters, with high-loss samples being targeted for the subsequent curation step. For the REMOVE approach, training samples with loss that are two standard deviations worse than the mean are dynamically removed during training, leading to the shrinking tail of the loss distribution. REPLACEIMG gradually reduces losses, resulting in the losses forming a mixture of Gaussians consisting of the original image-text pairs and the those with synthesized images. Going beyond
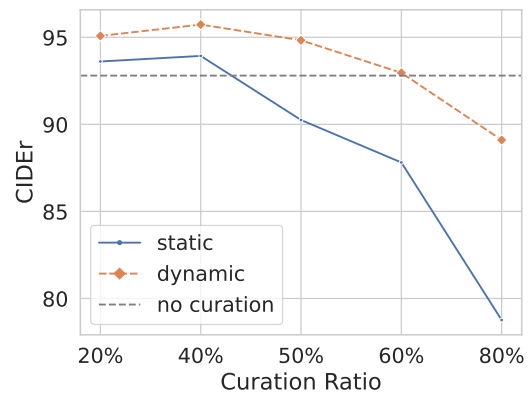
just the losses of the training samples, we also inspect the distributions of the words in the training captions for the curation methods. Figure 6 shows these distributions, where it can be seen that RE-MOVE reduces low-frequency and singleton words during training, while REPLACECAP increases the counts of some lower-frequency words while removing singletons. By definition, REPLACEIMG only changes the distribution of the images used to train the model, and as such, does not change the distribution of the words in the training data.

**The efficacy of dynamic replacement** Using training loss values as an effective indicator, we dynamically curate on the training samples identified as challenging. In REPLACEIMG, another static approach is to replace the identical images, i.e. $I_k$ in $\{(I_k, C_k^1), \ldots, (I_k, C_k^J)\}$, with unique synthesized images before training, instead of updating the training samples while training. With static image replacement, for each of the reference captions,
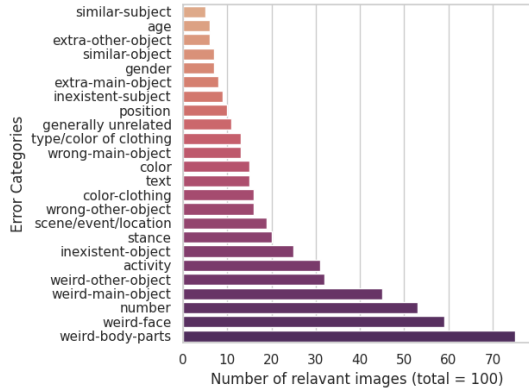
(a) Distribution of text-to-image generation errors.
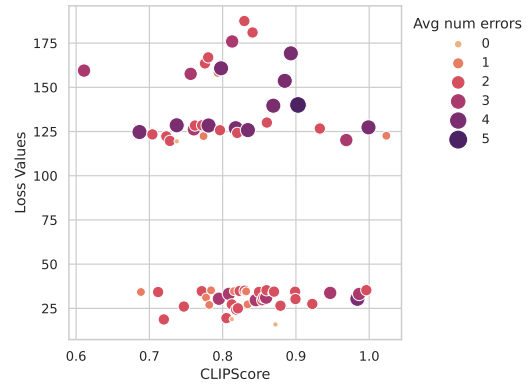


(b) Human evaluation versus CLIPScore.

Figure 8: Results of the human study of the errors made by the Stable Diffusion model in 100 images. The images used in the study were chosen to represent either low or high model loss. (a) Histogram of the number of errors annotated in each category. The most frequently occurring annotations concern weird deformations in the expected objects or humans. (b) Relationship between average number of identified errors by human annotations for each synthesized image and its captioning loss with regard to original captions. More errors are identified in images of higher loss. However, CLIPScore appears to fail in validating qualities of the synthesized images, as the score ranges are almost identical for samples that contain more errors.

we replace their original image with a generated image. Static replacement with 20%–80% curation ratio corresponds to replacing images for one–four captions of the original five. The 50% replacement ratio mimics a fair coin-flip, where for each of the text-image samples, there is 50% probability for the image to be replaced by a synthesized image.

We compare the efficacy of these two approaches in Figure 7. When evaluating on the original 1k validation set, we see that for both approaches, incorporating synthesized images of 20% or 40% can assist finetuning and achieves higher CIDEr scores. Nevertheless, dynamic image replacement consistently performs better than the static method, showing focusing on the hard samples is effective. For both replacement methods, performance starts to decrease when the curation ratio is too high. This may indicate that when incorporating too many images from the synthetic distribution, the gap increases between the training and evaluation sets.

**Replacing captions with LM generations**  As an alternative to the REPLACECAP method, we investigate the utility of replacing the captions with those generated by a language model (LM). Inspired by the approach in Ramos et al. (2023), we prompt the XGLM-2.9B model (Lin et al., 2022) with few-shot examples to generate a new caption. The LM generated caption is then paired with the image as the curated sample. We evaluate on Flickr30K using both models, applying the same curation ratio

| Method | BLIP | | BEiT3 | |
|---|---|---|---|---|
| | B | C | B | C |
| Baseline | 37.6 | 92.8 | 29.8 | 79.3 |
| +ReplaceCap | **37.9** | **94.5** | 29.6 | 80.1 |
| +ReplaceLMCap | 37.5 | 93.4 | **31.2** | **83.2** |

Table 2: Comparing caption replacement with LM generation to REPLACECAP on Flickr30K. Both methods improve over baseline for BLIP and BEiT-3.

of 1% as REPLACECAP. The results presented in Table 2 indicate that this approach can serve as a viable alternative to REPLACECAP, consistently outperforming baselines for both models. Please refer to Appendix A.3 for more implementation details.

**Human Study: Errors made by text-to-image generation models**  To assess the quality of the generated images and their alignment with human judgments, we perform a human study to evaluate the errors present in the synthesized images. This will serve to better understand any shortcomings with the REPLACEIMG curation that is not captured by automatic evaluation measures.

We first ranked synthesized images by model loss from the 1K images in the COCO validation set. We then sampled a subset for human annotation using the top and bottom 50 images based on their loss using our fine-tuned captioning model. These images are uniformly divided into 5 sets,

1080

each containing 20 images with equal number of the high loss ones and the low loss ones. The data was annotated by 12 people, members of a university research lab with a basic understanding of text-to-image generation but no knowledge of the bi-modal distribution of images. The annotators were asked to categorize the errors in the synthesized images, given both the image and the reference sentences that were used to generate the images. Each participant annotated one set images.

Starting from the categories defined by van Miltenburg and Elliott (2017), we defined 25 error categories including color, number mismatches, and errors related to people and objects in the images. Please see the user interface and more details in the Appendix A.1. We analyze the human judgements for the images that have at least three annotations, yielding 74 unique images.

As shown in Figure 8a, the most common problem of the synthesized images are that they often generate weird face or body parts, which makes the images less natural or pleasant. The text-to-image generation model is also weak at generating the correct number of people or objects. From Figure 8b we confirm the quality of our collected annotations that high loss figures often contain more errors on average. Furthermore, we note that CLIPScore is insensitive to these types of errors, indicating its limited capability of evaluating quality of generated images. Additional examples can be found in Figure 13 in the Appendix.

## 7 Further Analysis

With the human study revealing the failure modes of the text-to-image model, we now provide insights on various techniques that are proved useful for improving image relevance in curating the image captioning datasets.

**Round-trip captioning evaluation**    Most previous work in text-to-image generation uses image-oriented measures like FID (Heusel et al., 2017) or CLIPScore (Hessel et al., 2021). However, these measures are not suitable for our purpose as they are claimed to lack alignment with perceptual quality (Saharia et al., 2022). We also found that CLIPScore cannot distinguish between low- and high-loss samples in captioning (Figure 8).

Alternatively, similar to Hong et al. (2018), we use a fixed model to generate captions for synthesized images and then compare them to original captions in a three-step process (Figure 9): (1) Gen-
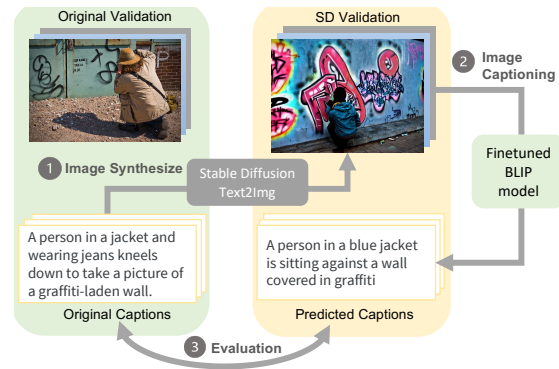


Figure 9: Round-trip captioning evaluation.

| Model | FT | Prompt | B | C | M |
|---|---|---|---|---|---|
| Upper-bound | | | 37.6 | 27.2 | 57.1 |
| SD 1.5 | - | concat | 31.0 | 24.7 | 52.5 |
| SD 1.5 | - | + styler | 30.8 | 24.2 | 52.5 |
| SD 1.5 | F | + styler | **33.5** | **25.0** | **53.5** |
| SD 1.5 | F | SBERT + styler | 30.6 | 24.1 | 52.0 |
| SD 2.0 | - | concat + styler | 31.2 | 24.8 | 52.0 |

Table 3: Round-trip captioning evaluation on Flickr30K with different Stable Diffusion models, prompts, and fine-tuning. F indicates that the model is finetuned. We report **B**LEU, **C**IDEr, **M**eteor.

erating images from validation set captions; (2) Predicting captions for the generated images using a strong image-captioning model; here we use BLIP fine-tuned on the COCO dataset but any other strong captioning model could be used instead. (3) Comparing the predicted captions with the original captions. The assumption is that if the generated images are of similar quality to the originals, the resulting captions will also be similar.

**Ablation on text-to-image variants**    Evaluating with round-trip captioning, we conduct an ablation study on variants of text-to-image generation models. Table 3 summarizes the evaluation results on the Flickr30K dataset. Specifically, we experiment with different versions of the Stable Diffusion models; prompt the diffusion models with various approaches (Section 3.2); and compare the generation performance between the finetuned text-to-image model and the pretrained ones. The results show that Stable Diffusion v1.5 finetuned on COCO outperforms the other variants, when prompted with the concatenation of all five captions, with the addition of the styler. For the details of the model variants, please refer to Appendix A.2.

## 8 Conclusion

In this paper, we have shown a simple, yet effective, data curation framework that can improve the performance of image captioning models. We investigated three approaches to data curation that dynamically update the training dataset based on high-loss image-caption samples. The methods involved either removing a sample, replacing the caption in a sample, or generating a new image from existing captions. Experimental results on the Flickr30K and MS COCO datasets show the effectiveness of these approaches to data curation without increasing the total size of the training dataset. A deeper analysis of the images synthesized by the text-to-image model shows frequent errors on generating objects of a certain amount or color, and struggles with human body features. A human evaluation of the errors in those images shows a clear difference in images with high or low losses.

In the future, we expect that better text-to-image generation models will lead to further improvements from using synthesized images to train image captioning models. From our insights in Appendix A.4, there is also significant promise on building a hybrid model combining different curation methods. We believe that a more sophisticated learning scheme leveraging multiple methods will offer more flexibility when curating the dataset. We plan on verifying whether these findings extend to other image captioning models. Moreover, we are also interested in applying the same framework to other multimodal tasks, especially those with under-complete datasets that cannot comprehensively cover the distributional space due to the cost of crowd-sourcing enough data, e.g. visual question answering, or visually-grounded dialog.

## Limitations

As Nguyen et al. (2023) has successfully improved the quality of the pretraining dataset by using an state-of-the-art BLIP-2 model to generate better captions, we would expect that our curation strategies to be scaled and adapted also to vision-language pretraining, which however is limited by research resources and therefore not explored in the scope of this paper. Currently our data curation methods also rely on state-of-the art pretrained models for both image understanding and text-to-image generation.

In our study, we explore how the application of various curation approaches impacts the down-stream image captioning performance under different curation ratios. While we predefine the curation ratio for our experiments in this paper, it is desirable for curation methods to be more readily applicable if the curation ratio can be automatically determined.

Moreover, while we take an online approach to data curation, our current approach is upper bounded in speed and performance of the text-to-image generation model. This might be a large bottle neck for adapting the strategy for more complicated vision-and-language tasks.

## Ethics Statement

Text-to-image generation is controversial in the broader AI and ethics community(Carlini et al., 2023). For example, it can generate images according to gender or racial stereotypes, which may prove harmful to members of those communities (Li et al., 2022b). While have not yet been observed in the vision-language domain, Shumailov et al. (2023) provide evidence that the use of synthetic data from generative models like large language models can introduce a potential risk of data quality degradation.

In this paper, we use text-to-image to improve the quality of an image captioning model, given a specific set of crowd-sourced captions. Those captions may themselves contain harmful stereotypes that would become more prevalent in our dynamically updated training datasets. As we dynamically update the model with new images based on loss values, we remove the water-marker in our generated images to prevent information leak to the model. Use of the synthesized images will strictly follow community guidelines.

While developing our curation methods that involve text-to-image generation for image replacement, we employed the stable-diffusion v1.5 model (Rombach et al., 2022), which was trained on the LAION-5B dataset. We note that we were unaware of any investigation into illegal material in the dataset (Thiel, 2023). Hence, we emphasize that our proposed framework is compatible with any other text-to-image models trained on more reliable datasets. Taking this in to consideration, we encourage researchers to explore and apply alternative text-to-image models when incorporating the curation techniques in their future work.

## Acknowledgement

## References

Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. 2021. A course-focused dual curriculum for image captioning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*.

Viktar Atliha and Dmitrij Šešok. 2020. Text augmentation using bert for image captioning. *Applied Sciences*.

Hammad Ayyubi, Rahul Lokesh, Alireza Zareian, Bo Wu, and Shih-Fu Chang. 2023. Learning from children: Improving image-caption pretraining via curriculum. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. 2023. Synthetic data from diffusion models improves imagenet classification.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2).

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Davide Caffagni, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Synthcap: Augmenting transformers with synthetic data for image captioning. In *International Conference on Image Analysis and Processing*, pages 112–123. Springer.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144.

Zui Chen, Lei Cao, and Sam Madden. 2023. Lingua manga: A generic large language model centric system for data curation. *arXiv preprint arXiv:2306.11702*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. 2021. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. 2023. Distilling model failures as directions in latent space. In *The Eleventh International Conference on Learning Representations*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for*

*Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Minghui Li, Yan Wan, and Jinping Gao. 2022b. What drives the ethical acceptance of deep synthesis applications? a fuzzy set qualitative comparative analysis. *Computers in Human Behavior*, 133:107286.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Fenglin Liu, Shen Ge, and Xian Wu. 2021. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online. Association for Computational Linguistics.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*, volume 139, pages 8162–8171. PMLR.

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.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In

*Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers. 2021. Changing the world by changing the data. In *Annual Meeting of the Association for Computational Linguistics*.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

David Thiel. 2023. Investigation finds ai image generation models trained on child abuse.

Emiel van Miltenburg and Desmond Elliott. 2017. Room for improvement in automatic image description: an error analysis. *CoRR*, abs/1704.04198.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Changrong Xiao, Sean Xin Xu, and Kunpeng Zhang. 2023. Multimodal data augmentation for image captioning using diffusion models. *arXiv preprint arXiv:2305.01855*.

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. *TACL*, 2:67–78.

Hongkuan Zhang, Saku Sugawara, Akiko Aizawa, Lei Zhou, Ryohei Sasano, and Koichi Takeda. 2022. Cross-modal similarity-based curriculum learning for image captioning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

## A  Appendix

### A.1  User interface for human study on categorizing text-to-image generation errors

Our user interface is shown in Figure 10. Annotators were asked to tick boxes of errors that they found in the given synthesized images.

The error categories include:

- **People**: age, gender, type of clothing, color of clothing, weird face, weird body

- **Main object**: wrong, similar, inexistent, extra, weird

- **Other objects**: wrong, similar, inexistent, extra, weird

- **General**: stance, activity, position, number, inconsistent references, scene/event/location, text, color, generally unrelated
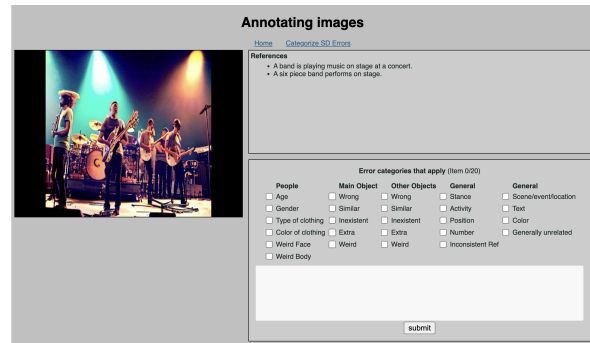


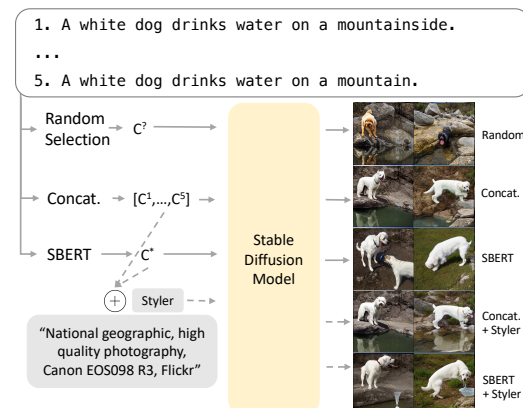Figure 10: Annotation interface for categorizing SD errors.



Figure 11: Different prompting strategies for synthetic image generation with text-to-image generation and representative examples. Based on our Round-trip Captioning Evaluation, prompting with the concatenated captions and the styler generates the best images for the task.

### A.2  Prompting approaches for text-to-image generation

Figure 11 illustrates the different approaches that we use to prompt the text-to-image generation model. We manually design the styler by inspecting the generated visual examples.

### A.3  Generating alternative captions with XGLM

We follow the prompt template used in (Ramos et al., 2023) to obtain LM-generated captions, i.e. "I am an intelligent image captioning bot. Similar images have the following captions: <captions> A creative short caption I can generate to describe this image is: <generation>". Here we used four ground truth captions as <captions> and the other one in <generation> for a image to build three-shot examples as the prompt. We used the 'facebook/xglm-2.9B' model which is available on HuggingFace (Wolf et al., 2019). We set the maxi-

mum generation length to 30 tokens with number of beams of 5 to prevent from generating repeated tokens.

## A.4 Combining multiple curation methods

In our pursuit to assess the efficacy of a hybrid model incorporating multiple curation methods, we experiment on the Flickr30K dataset with BEiT-3 as an initial attempt. For the combining strategy, we selected the two most effective methods on the dataset, namely REMOVE and REPLACEIMG. After each training epoch, we curated the training samples by eliminating one half of the top loss samples while substituting the images of the remaining half. Here we curate on the samples with a loss that exceeded two standard deviations from the mean. Our experiment achieves a CIDEr score of 83.8 and a BLEU4 score of 32.8, surpassing previous single curation performance on the dataset. We believe that the hybrid curation approach would yield greater benefits with more sophisticated combining strategies, which we leave for future work.

## A.5 High-loss training samples

In Figure 12, we visualize the high loss training samples in the COCO dataset after the first epoch of finetuning. These samples are target of our curation techniques. Compared to the average caption length of 11 words, the top samples all have very long captions of around 30 words, making it difficult for the model to learn. In the following finetuning epochs, we curate on these samples by either removing the text-image pairs completely (REMOVE), replacing the caption (REPLACECAP), or replacing the image with a synthesized unseen image (REPLACEIMG).

## A.6 Examples of synthesized images

In Figure 13, we show examples of synthesized images from the text-to-image model that are of high losses and low losses, alongside with the human annotations regarding errors identified from these images.

| Image | Caption | Length | Loss |
|---|---|---|---|
| | a picture of a clearly disrespectful person littered, abused alcohol, didn't flush their bad choices, and worst of all, let old glory touch a bathroom floor | 26 | 213.24 |
| | a picture of a rain-wet street view with lots of bike riders, rimmed with buildings that seem to bunch up and fight for space might look gray and unprepossessing, but doesn't, in part | 33 | 200.14 |
| | a picture of the scene shows outdoors, furthest to closest, shrubbery than a playing field with at least two uniformed and young players, and closest, a blue fence, and a long bench with | 33 | 200.02 |
| | a picture of it is outdoors, the exterior of a low roofed domicile, where a tiny grove of slender tropical trees makes a lean-to for super-modern blue and white motorcycle | 30 | 199.90 |
| | a picture of while a purple/blue sky with what looks like a kite or a loose para-sail floating in it covers most of a distance shot, the bottommost part shows grassy side banks | 33 | 197.36 |

Figure 12: High loss training samples in COCO after the first epoch, ranked by loss in descending order. The top samples all have very long captions around 30 words, compared to the mean of 11 words of the datasets.

| Image | Caption | CLIPScore | Loss | Categorized Errors |
|---|---|---|---|---|
|  | A picture of two women with one in lacy white dress with handbag and leggings and the other with a tall red hat, black mid-dress, and frame like plastic dress on top. | 84.1 | 181.0 | type/color of clothing, color-clothing, weird-face |
|  | A pedicab driver waiting on his bike. | 89.3 | 169.2 | weird-main-object, weird-other-object, weird-body-parts, stance |
|  | A man in a black suit with tie and corsage smiles at a girl who smiles back, both are sitting at a table at a semi formal event such as a wedding or reunion. | 77.6 | 163.5 | color-clothing, weird-body-parts, wrong-main-object, scene/event/location |
|  | Two men are playing guitars and one man is singing into a microphone on a stage with the spotlight on them. | 74.7 | 26.0 | weird-face, weird-body-parts, weird-main-object, weird-other-object |
|  | There a several people in a dark bar-type room, including one girl on a stool. | 84.9 | 26.5 | number, weird-face, weird-main-object, weird-body-parts |
|  | Many children are playing and swimming in the water. | 78.2 | 26.9 | weird-face, weird-body-parts |

Figure 13: Examples of synthesized images that are of high losses (top) and examples of synthesized images that are of low losses (bottom). Human annotations show that consistent error types have been recognized for the high loss samples while CLIPScore fails to align with human judgement. The low loss synthesized images are visually less complicated than the higher loss ones, but can still often look weird and contain errors in color or objects.