Targeted Image Data Augmentation Increases Basic Skills Captioning Robustness

Valentin Barriere[1,2]*, Felipe del Rio[1,3]*, Andres Carvallo de Ferari[2]*, Carlos Aspillaga,[1] Eugenio Herrera-Berg,[1] Cristian Buc[1]
[1]Centro Nacional de Inteligencia Artificial, Macul, Chile
[2]Department of Computer Science, Universidad de Chile, Santiago, Chile
[2]Department of Computer Science, Pontificia Universidad Catolica, Santiago, Chile
name.lastname@cenia.cl

Abstract

Artificial neural networks typically struggle in generalizing to out-of-context examples. One reason for this limitation is caused by having datasets that incorporate only partial information regarding the potential correlational structure of the world. In this work, we propose TIDA (Targeted Image-editing Data Augmentation), a targeted data augmentation method focused on improving models' human-like abilities (e.g., gender recognition) by filling the correlational structure gap using a text-to-image generative model. More specifically, TIDA identifies specific skills in captions describing images (e.g., the presence of a specific gender in the image), changes the caption (e.g., "woman" to "man"), and then uses a text-to-image model to edit the image in order to match the novel caption (e.g., uniquely changing a woman to a man while maintaining the context identical). Based on the Flickr30K benchmark, we show that, compared with the original data set, a TIDA-enhanced dataset related to gender, color, and counting abilities induces better performance in several image captioning metrics. Furthermore, on top of relying on the classical BLEU metric, we conduct a fine-grained analysis of the improvements of our models against the baseline in different ways. We compared text-to-image generative models and found different behaviors of the image captioning models in terms of encoding visual encoding and textual decoding.[1]

1 Introduction

Humans and animals develop all kinds of cognitive abilities from a very early age that allow them to interact with their world (Spelke et al., 1992; Spelke and Kinzler, 2007). For instance, infants display numerical cognition abilities (Feigenson et al., 2004; Xu and Spelke, 2000), can recognize emotions (Bornstein and Arterberry, 2003) or even the danger associated with other agents' action plans (Liu et al., 2022a). Comparatively, animals also display similar numerical cognition abilities (Davis and Memmott, 1982; Dacke and Srinivasan, 2008), or recognize emotions in order to better communicate within a social group (Hantke et al., 2018). These abilities are crucial in order to build models of the world that are necessary for planning, reasoning, and solving complex decision-making tasks (Lake et al., 2017).

Deep learning systems can solve these tasks by optimizing an objective function via supervised, semi-supervised or unsupervised learning (LeCun et al., 2015). Within this framework, it has been shown that deeper layers progressively represent increasingly abstract concepts (Krizhevsky et al., 2017), akin to what has been observed in the human visual or auditive processing pathways (Cichy et al., 2016; Caucheteux et al., 2023). Moreover, empirical work has shown that pretrained state-of-the-art transformer models (Devlin et al., 2019) encode factual knowledge within sets of knowledge neurons (Dai et al., 2022); strongly related to the concepts of "grandmother" cells in neuroscience (Quiroga et al., 2005). Importantly, not only factual knowledge but also conceptual knowledge (such as "sentiment" in a text or "written language" in an image) are encoded by nodes in deep layers (Radford et al., 2017; Yosinski et al., 2015). Whereas recent methods have been proposed to access and edit factual knowledge (Meng et al., 2022b), and thus evaluate how and where facts are being encoded in deep networks (Meng et al., 2022a), it is much harder to evaluate the abilities associated with conceptual knowledge stored in these networks. Yet, possessing such a conceptual knowledge base is crucial for out-of-distribution generalization (Bosselut et al., 2019).

Although deep networks seem to encode conceptual knowledge that allows them to display human-like abilities such as counting, emotion, gender,

[1]Code will be available online after submission.

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 243–257

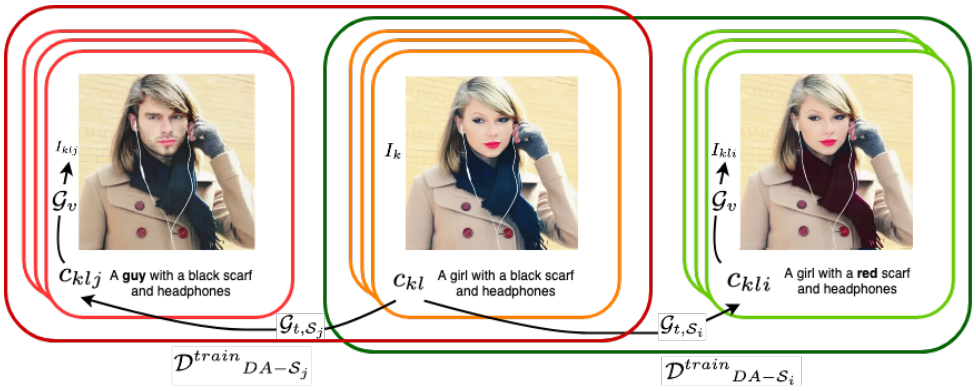

Figure 1: TIDA Framework (Example generated with Null-Text-Inversion (Mokady et al., 2022))

color, and sentiment recognition/categorization (Wallace et al., 2019; Barriere et al., 2022; Hendricks et al., 2018; Anderson et al., 2016a; Barriere, 2017), these same networks typically struggle in producing out-of-context (or out-of-distribution) generalizations (Marcus, 2018; Lake and Baroni, 2018; Ruis et al., 2020; del Rio et al., 2023; Ribeiro et al., 2020). These limitations are due to the inherent functioning of Artificial Neural Networks (ANNs). Indeed, generalization performances of ANNs largely depend on their ability to extract the correlational structure in the training data set, memorize this structure, and extrapolate it to a novel (test) data set (Krizhevsky et al., 2017; Saxe et al., 2019).

Indeed, given that the performance of vanilla deep networks is constrained by the structural correlation observed in the training data set, a straightforward way to maximize the generalization performance in ANNs is to augment data sets in *targeted* ways (Sharmanska et al., 2020; He et al., 2023). Thereby, targeted data augmentation would increase the span of potential correlations that could be observed in the world, and as such improve the human-like abilities of deep networks. By targeting specific human-like abilities and augmenting the data set to encapsulate unseen examples associated with these abilities, we hypothesize that models can increase their conceptual knowledge, and thus improve their performance on specific benchmarks we discuss below. Moreover, similar to editing unique factual knowledge (Meng et al., 2022b), one would ideally want to target unique conceptual knowledge (e.g., gender, color, numerosity, emotion, shape...) to induce such ability-selective performance, which has been widely studied (Anderson et al., 2016b; Hu et al., 2023).

We will propose a simple way to overcome the issues raised above, for Image Captioning (IC) task. Interestingly, novel text-to-image generation models (Rombach et al., 2022b) in combination with text-generation or manipulation (He et al., 2023; Mitkov, 2022; Murty et al., 2022) affords novel possibilities for targeted data augmentation for vision-language tasks. Hence, we propose to enhance the capabilities of an Image Captioning model by using a targeted data-augmentation on several specific abilities (or skills). We use simple regular expressions (regex) to identify these skills in the caption, to change the caption for another version of it, and to generate the image related to this caption. The main contributions of this work are twofold. First, we propose a simple method to identify data related to a specific human-like ability in image captioning (e.g., color identification, emotion recognition...). Second, we propose a novel data augmentation method based on image-to-text generation models that allows one to generate data sets that can selectively improve a single or combinations of human-like skills in image captioning performance. Instead of manipulating or fine-tuning information processing within image captioning models, our method increases the span of potential object correlations and thus allows us to generalize image captioning abilities to a broader spectrum of situations that can be observed in the real world (Zhang et al., 2021). In what follows, we first describe related work while specifying the original contribution of our work. Subsequently, we describe the Targeted Image-editing Data Augmentation (TIDA; see Figure 1) method and present the results associated with fine-tuning models with our TIDA-augmented data sets. Finally, we discuss the implications of our work.

2 Previous and Related Work

Image Captioning Image captioning (IC) models provide human-like captions to images (Cornia et al., 2020; Herdade et al., 2019). Such an ability lies in the intersection between computer vision and natural language processing (Devlin et al., 2015), and is therefore, in essence, a multimodal problem. Early IC models proposed to sequentially combine convolutional neural networks (CNN) with recurrent neural networks (RNN) into a single imaged-conditioned language model (Karpathy and Fei-Fei, 2015; Chen and Lawrence Zitnick, 2015; Fang et al., 2015). Given the success of these models and their potential industrial applications, subsequent work has focused on improving the models' image captioning ability by focusing on specific properties of IC models. For instance, it has been shown that top-down visual attention mechanisms improve captioning performance (Anderson et al., 2018; Lu et al., 2017). Alternatively, focusing on the learning process, it has been shown that implementing self-critical sequence training (a variant of the REINFORCE algorithm) improves IC performances by avoiding the exposure bias (Ranzato et al., 2016) and directly optimizing the relevant task metrics (Rennie et al., 2017). Furthermore, many IC models are pre-trained using tasks like Masked Language Modeling (MLM) and Image-Text Matching (ITM). These tasks possess losses that differ from image captioning (or other downstream tasks), and thus IC models require further fine-tuning. Hence, recent work has focused on unifying generative vision-language models through text generation (Cho et al., 2021; Wang et al., 2022a,b), in order to optimize knowledge transfer from train to test. Lastly, novel methods have focused on optimally leveraging language caption supervision during pre-training, as small datasets with large caption variability can lead to detrimental effects (Santurkar et al., 2023).

Symbolic Knowledge Vision-language (VL) tasks can also be improved by incorporating symbolic knowledge into the VL models. For instance, providing a knowledge base, instantiated as subject-relation-object triplets associated with the images, both improve performance in vision-question answering (VQA) tasks, on top of allowing to explain the VQA model's predictions (Riquelme et al., 2020). In the same vein, adding high-level (semantic) attributes as inputs to IC models can increase captioning benchmarks (You et al., 2016; Yao et al., 2017). Alternative efforts have shown that using object tags to facilitate the semantic image-text alignment during pre-training, and improves benchmark metrics in downstream fine-tuned image captioning tasks (Li et al., 2020). Moreover, aligning directional semantic and spatial relationships between text and image (i.e., relation-level alignment) improves compositional reasoning (Pandey et al., 2022). Finally, symbolic knowledge and reasoning capability aim to enhance textual model's robustness when faced with out-of-distribution examples, thereby enabling them to engage in more human-like reasoning (Collins et al., 2022).

Bias/Bug detection, and Evaluation TIDA enhances the likelihood of simultaneously observing distinct attributes in an image within the augmented dataset. Thereby, our work relates to studies that focus on improving the predictive abilities of models in domains that suffer from bias-induced incorrect predictions. In line with this idea, the *Equalizer* model is constrained to attend to the person attribute in images, increasing the IC abilities to detect the gender in the image (Hendricks et al., 2018). Interestingly, other attributes such as numeracy (e.g., counting) naturally emerge in standard embeddings (Wallace et al., 2019), and may thus be less prone to biased predictions. Alternative debiasing methods focus on "decoupling" biased directions within text embeddings (Chuang et al., 2023).

Other approaches focus on discovering the specific images where IC models fail (i.e., bugs). An instance of such a method uses a sequential pipeline that generates images from specific captions, classifies the object in the image, creates captions from the incorrectly classified images, generates captions of these images, and finally regenerates novel images based on the previously generated caption via a text-to-image generative process. These last images can be used to assess the robustness of vision models, as well as improve their performance (Wiles et al., 2022).

Moreover, while image captioning is usually scored on automatic metrics like SPICE (Anderson et al., 2016b) or CIDEr (Vedantam et al., 2015), it has been suggested that metrics evaluating both precision *and* recall leading to better correlations with human judgments (Kasai et al., 2022). Finally, (Hu et al., 2023) propose a method to compare image captioning models correlated with human

judgment by leveraging LLM (OpenAI, 2023).

Data augmentation and Image generation Data augmentation has been shown to improve performance both in vision (Ho et al., 2019; Cubuk et al., 2020) and language (Sennrich et al., 2015; Karimi et al., 2021; Andreas, 2020; Wei and Zou, 2019) tasks. Typically, data augmentation techniques involve procedures such as geometric transformations, color space augmentations, kernel filters, or mixing images (see (Shorten and Khoshgoftaar, 2019) for review). To further improve these augmentation methods, a multi-task view of augmentation proposes to incorporate both original data and augmented examples during the training procedure (Wei et al., 2021). This proposal has the benefit to relax the assumption that augmented examples cannot be too dissimilar from the original data. In the same vein, *Neurocounterfactuals* is a method that allows augmenting data via large counterfactual perturbations that still bear resemblance to the original data but can nonetheless provide richer data augmentation (Howard et al., 2022). More recent studies have investigated data augmentation methods in multimodal settings such as VL tasks. For instance, LeMDA is a method that learns an augmentation network alongside a task-dedicated network (Liu et al., 2022b). This method augments the latent representation of the network and thus preserves the semantic structure in each modality.

Moreover, not restricting data augmentation to the specificity of inputs can have detrimental effects, as augmented examples may possibly be associated to another label (e.g., a color change from green to red rock may induce a label change from emerald to ruby). To avoid this pitfall, instance-specific augmentation (*InstaAug*) learns to apply invariances to specific parts of the input space (Miao et al., 2022). Similar work suggests estimating invariances by learning a distribution over augmentations, and jointly optimizing both the network *and* augmentation distribution parameters (Benton et al., 2020).

Other methods belong to a class of automated data augmentation algorithms. These algorithms can for example use reinforcement learning (RL) to optimize a data augmentation policy (e.g., (Cubuk et al., 2019)). Furthermore, differentiable data augmentation proposes a method that relaxes the discrete state search assumption of RL, and allows for a more efficient data augmentation by implementing an end-to-end differentiable search procedure (Hataya et al., 2020). Notably, other methods such as *AdaAug* extend previous research by focusing not only on instance-depend data augmentation but also on class-dependent ones through the implementation of adaptive augmentation policies (Cheung and Yeung, 2022).

Our method differentiates from policy-based methods for data augmentation but remains both automated, class-dependent, and targeted (i.e., we can focus on specific attributes such as gender, counting, or color). In particular, we leverage the impressive natural language-driven image synthesis abilities of text-to-image generative models (Yu et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) (see methods). In particular, we focus on their image editing or inpainting ability, which is a difficult challenge for these models given that only part of the image has to be changed while the rest has to be maintained. To solve this issue, traditional methods make use of explicit masks to circumscribe the inpainting region (Nichol et al., 2022; Avrahami et al., 2022). However, masking methods are both time-consuming and do not leverage structural information in the image. To circumvent this issue, recent work proposes the use of a prompt-to-prompt procedure in combination with a cross-attentional control mechanism that allows to edit of specific objects in the image while taking into account the contextual information (Hertz et al., 2022). Another method proposes to use of null-text inversion to achieve maskless image editing (Mokady et al., 2022).

Interestingly, these state-of-the-art inpainting models open up the possibility to implement novel data augmentation methods. For instance, a recent paper showed that fine-tuning large-scale image-to-text generative models allows producing high-quality synthetic data that can improve ImageNet benchmark scores (Azizi et al., 2023). TIDA extends this idea in VL models, in order to improve specific target skills of these models within the framework of image captioning tasks.

3 Method and Experiments

We propose a two-step method that allows retrieving certain images using their captions, regarding a specific concept that we call *skill*. These skills refer to human- and animal-like basic abilities, such as gender categorization, counting, or recognizing colors. We first use a text mining method to

detect whether or not a caption contains specific words that are related to the skill (Subsection 3.1). Second, we generate variants of the original skill-related captions and create new images with these new captions in order to augment the dataset for each type of skill (Subsection 3.2). An overview of the method is shown in Figure 1.

3.1 Skill-related retrieval

We assume a list of S skills $\{\mathcal{S}_i, i = 1...S\}$, a training dataset of captions and images $\mathcal{D}^{\text{train}} = \{(\mathbf{C}_k, I_k), k = 1..k_{\text{train}}\}$, $\mathbf{C}_k$ being a set of ground truth captions.

For each skill $\mathcal{S}_i$ we create a binary classifier $f_{\mathcal{S}_i}$ that detects whether or not the skill $\mathcal{S}_i$ is present in a pair of image and associated captions. By applying this function to a dataset $\mathcal{D}$, it is possible to create a subpart of this dataset $\mathcal{D}_{\mathcal{S}_i}$ containing samples related to the aforementioned skill. By using this method and for each skill $\mathcal{S}_i$, we retrieve a subpart of the train $\mathcal{D}^{\text{train}}$ dataset that we call ${\mathcal{D}^{\text{train}}}_{\mathcal{S}_i}$ and a subpart of the test $\mathcal{D}^{\text{test}}$ dataset that we call ${\mathcal{D}^{\text{test}}}_{\mathcal{S}_i}$. The former will be used for data-augmentation and the latter will be used for the evaluation of the different models.

3.2 Targeted Data Augmentation

In order to improve the performances of the model with regard to several skills, we augment the dataset with sets of new examples. Those examples are created so that they depict new situations that are not necessarily in the training set, but should help the model generalize. For this purpose, we create a set of text generators functions $\{\mathcal{G}_{t,\mathcal{S}_i}, i = 1...S\}$ taking as input a text caption containing a skill $\mathcal{S}_i$ and generating a slightly different version of this caption. The generator function perturbs the caption's text in such a way that it remains related to the skill. For example, it would inverse the gender of one of the words in the sentence: The caption "a man is playing basketball" would be changed (or perturbed) to "a woman is playing basketball". Mathematically, for any caption c_{kl}[2] containing the skill $\mathcal{S}_i$, we create another caption $c_{kli} = \mathcal{G}_{t,\mathcal{S}_i}(c_{kl})$.

Finally, for every perturbated caption c_{kli} we use a text-to-image generator $\mathcal{G}_V$ in order to create an image I_{kli} associated with the novel caption. We obtain an artificial set of image-caption pairs, which gives with the original images, the dataset ${\mathcal{D}^{train}}_{\mathcal{G}_V - \mathcal{S}_i}$.

[2] caption l of the image k

Those augmented datasets ${\mathcal{D}^{train}}_{\mathcal{G}_V - \mathcal{S}_i}$ are used to train several image captioning models, which should focus more on the specific skill $\mathcal{S}_i$. Each of the models is then evaluated on the different test sets ${\mathcal{D}^{test}}_{\mathcal{S}_i}$ which contain the pairs of images and list of captions that are related to the skill $\mathcal{S}_i$. The pseudo-code is visible in Algorithm 1.

Algorithm 1 The TIDA method on train

Require: Skills $\mathcal{S}_i$, Textual skill detectors $f_{\mathcal{S}_i}$, Text generators $\mathcal{G}_{t,\mathcal{S}_i}$, Image generator $\mathcal{G}_V$, Train set $\mathcal{D}^{\text{train}} = \{(c_{kl}, I_k)\}$

for i in $1...S$ **do**

  ${\mathcal{D}^{\text{train}}}_{\mathcal{G}_V - \mathcal{S}_i} \leftarrow \mathcal{D}^{\text{train}}$ ▷ Initialize

  ${\mathcal{D}^{\text{train}}}_{\mathcal{S}_i} \leftarrow f_{\mathcal{S}_i}(\mathcal{D}^{\text{train}})$ ▷ IC pairs with skill i

  for (c'_{kl}, I'_k) in ${\mathcal{D}^{\text{train}}}_{\mathcal{S}_i}$ **do**

    $c'_{kli} \leftarrow \mathcal{G}_{t,\mathcal{S}_i}(c'_{kl})$ ▷ Caption perturbation

    $I'_{kli} \leftarrow \mathcal{G}_V(c'_{kli})$ ▷ Image generation

    ${\mathcal{D}^{\text{train}}}_{\mathcal{G}_V - \mathcal{S}_i} \leftarrow {\mathcal{D}^{\text{train}}}_{\mathcal{G}_V - \mathcal{S}_i} \cup \{(c'_{kli}, I'_{kli})\}$ ▷ Adding the new pair

  end for

end for

3.3 Dataset

For the image captioning task, we use the Flickr30K (Young et al., 2014), which is composed of 31K photographs of everyday activities, events, and scenes harvested from Flickr and 159K captions. Each image is described independently by five annotators who are not familiar with the specific entities and circumstances depicted in them. We follow Karpathy's split[3] (Karpathy and Fei-Fei, 2017), which gives 29.8k/1k/1k images for train/val/test.

3.4 Methodology

Skill used We augment the data regarding three basic human skills: gender detection, counting capability and color recognition. We focus on these skills for consistency with previous work (Anderson et al., 2016b), and because they are considered as essential and acquired early in humans and present in animals (Wang et al., 2010; Dacke and Srinivasan, 2008; Davis and Memmott, 1982).

Text generation For each skill, and for each of the captions that were retrieved as containing it, we

[3] cs.stanford.edu/people/karpathy/deepimagesent/captiondatasets.zip

Test / Train	#DA	BLEU@1-4				RefCLIPScore				Spice		
		$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$	$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$	$F1_{clr}$	$F1_{ctg}$	$F1_{all}$
$\mathcal{D}^{train}$ (Vanilla)	0	51.8	44.0	49.9	49.7	79.9	79.3	79.8	80.3	24.1	19.7	20.7
$\mathcal{D}^{train}{}_{SD-rnd}$	60k	51.3	44.1	49.2	49.6	80.0	79.5	79.7	80.2	**24.7**	**25.2**	20.6
$\mathcal{D}^{train}{}_{SD-clr}$	20k	*51.7*	44.0	*49.3*	49.5	79.8	79.4	79.6	80.1	24.3	19.8	20.2
$\mathcal{D}^{train}{}_{SD-ctg}$	20k	*51.7*	*44.4*	49.2	49.7	79.9	*79.5*	79.7	80.2	23.4	22.0	20.4
$\mathcal{D}^{train}{}_{SD-gdr}$	20k	51.2	43.4	48.5	48.8	*80.0*	79.2	*79.9*	80.3	24.5	24.4	20.6
$\mathcal{D}^{train}{}_{SD-all}$	60k	51.8	44.9	**50.1**	**50.5**	80.1	**79.7**	80.1	**80.5**	**24.7**	23.6	**21.0**

Table 1: Average of the BLEU@1-4 scores of the different TIDA-enhanced models on the different test sets. The TIDA models depicted used different image generation strategies: *SD* uses Stable Diffusion and *AAE* Attend-and-Excite. The first line contains the performance of the model trained with the Vanilla train set. Then, the first to third line of each TIDA model contain the results of the model trained with data-augmentation on the color, counting, and gender skills, respectively. And, the last line of each, depicts the results of the model trained with all three types of data-augmentation. The scores in bold are the best scores on each test set, while the scores in italic are the best scores of each of the models trained with (skill-related) data-augmentation.

changed the caption text by using an alternative attribute of the targeted skill. For this, we employed a list of defined words that were related to the targeted skills. Each of the skill-related words has a list of other words that can be used as a replacement. For gender, masculine words like "man" were replaced by their feminine counterparts like "woman". For color, we swapped the different colors altogether. For counting, we either added or subtracted 1 to the detected written number in the sentence (± 1). See Appendix A for more details.

Baseline We compared our method with a data-augmentation that consists of generating images from random captions of the dataset. In this way, we aim to show that the improvement in different performances do not only come from having a larger training set, but also to have a larger and more diverse training set. In the following, we call this augmented training set $\mathcal{D}^{train}{}_{SD-rnd}$.

3.5 Implementation details

Text generator We used simple regular expressions to find the different attributes of each skill. The replacement words were chosen randomly within the list of possible alternatives. More details are available in Appendix A.[4]

Image generator We test a classical text-to-image generation technique with Stable Diffusion (Rombach et al., 2022b) and generated 20k images per skill. For Stable Diffusion, we used the version 1.5[5] as described in (Rombach et al., 2022b), leveraging the Diffusers library for its implementation (von Platen et al., 2022). We used a 16-bit floating-point data type and a guidance scale set at 8, which constrained the extent to which textual prompts generated the resultant images. The resolution of the generated images was 128 x 128 pixels. The remainder of the parameters were set as default, as specified by the Diffusers library. In the Appendix B, we show experiments with more generators.

Image captioning We used the BLIP model (Li et al., 2022) because of its state-of-the-art performances on Image Captioning, with a publicly available code and pre-trained weights. We kept the same original hyper-parameters, adjusting only the batch size from 32 to 24 and using the ViT Base model as the image encoder, due to hardware limitations. For the training, we also kept the AdamW (Loshchilov and Hutter, 2019) original optimization algorithm with an initial learning rate of 10^{-5} that is decreased through the training based on a $cos(\cdot)$ function until it reaches 0. In order to compare models with different amounts of available data, we used early stopping with a patience of 5.

Metrics We used the classical BLEU metric (Papineni et al., 2002) to evaluate the performances of the models. Moreover, we used another metric that relies on learned representations. We computed RefCLIPScore (Hessel et al., 2021) which is based on the similarity between the embedding of the caption and the embedding of the image coming from CLIP (Radford et al., 2021). This metric was shown to have a better correlation with human judgments than other classical metrics (Kasai et al., 2022).

[4] All our code will be made available after publication.

[5] https://huggingface.co/runwayml/stable-diffusion-v1-5

4 Results and Analysis

4.1 Results

The results of the models trained with different skill-based data-augmentation on different test sets are shown in Table 1. We can see that the overall best scores on each test set are obtained with the model using the three types of data-augmentation techniques, either using BLEU (from 49.7 to 50.5) or RefCLIPScore (from 80.3 to 80.5).

We also provide the F1-scores computed with Spice, and especially the ones related to counting and color because we aim to quantify the performances of the models on those skills. The data-augmentation helps to augment both of the metrics individually, more than the overall one.

4.2 Analysis

We analyze the results in three different ways: (i) by using classical natural language generation metrics for image captioning, (ii) by assessing the use of skill words regarding the captions and quantifying the right use of the skill-related terms, (iii) by probing the representation of the image on a skill detection task for a finer comprehension of the image encoder and text decoder behavior.

Classical metrics By analyzing the classical metrics we can make several observations. Contrary to what we would have expected, the skill-related TIDA are not necessarily leading to the best scores in their respective test sets. Note however that the metrics are not homogeneous. The counting-related TIDA obtains the best results on the counting test set for BLEU and RefCLIPScore, but Spice F1-counting is better with gender. Interestingly, counting (compared with color and gender) leads to the worst metrics with BLEU but the best one when focusing on the RefCLIPScore and Spice metrics. More details and metrics are available in Appendix C.

Skill-related words In order to analyze the results of the model by going beyond the classical opaque metrics like BLEU and RefCLIPScore, we used a similar method to spice (Anderson et al., 2016b) that allows to investigate specific semantic words. TIDA relies on using certain variations of words, hence we are evaluating the propensity of the model to use those words in the right context. If a word associated with a skill is present in the ground truth or in the generated caption, it allows us to quantify the results of the model as false/true positive/negative. Specifically, when the model is using a word associated with a skill in the generated caption, and this skill is indeed associated with the image-caption ground truth, we count this as a true positive. If the model does not use any word associated with a skill and the skill is not present in the ground truth, we count this as a true negative. The other combinations are regarded as false positives or negatives. The precision, recall, and F1 for color, counting, and gender TIDAs are available in Table 2.

For the color TIDA, the precision and recall are both increasing for the positive and negative cases. This means that the model is using more often color words when the caption should contain one and less when it should not. For the counting TIDA, the recall of the negative class is augmenting from 39.1 to 45.9, which means that the model uses fewer counting-related words when it should not. At the same time the precision for the positive class augments which means the use of counting-related words is more pertinent. For the gender TIDA, the model is using more gender words (recall positive going from 88.8 to 92.4) while being a bit less precise (recall negative decreasing from 79.0 to 77.8). Overall, we observe that the color TIDA gives better results for color, but surprisingly the counting TIDA is better for gender and the gender TIDA is better for counting.

Probing with visual representations We tried to analyze how TIDA influences the model not only using the raw results of the text decoder but also using the representation of the image encoder. For this purpose, we proposed to probe the image representations to predict whether or not the image is associated with a specific skill.

As we previously did, we used the text-mining method to label whether or not a sample is associated with one of the three skills. We then trained a linear multi-layer perceptron on the representations produced by the image encoder and these labels. As is usual with transformer-based models, we used the class embedding coming from the image encoder as the image representation embedding. We use binary cross entropy loss and SGD to train the probe and perform early stopping and a grid search on each model to find the best model hidden size and learning rate. The results with the five TIDA models are shown in Table 3.

Looking at the F1-score, it seems that none of the TIDAs bring any significant change regarding

Skill	Color					Counting					Gender				
Train	P+	R+	P-	R-	F1	P+	R+	P-	R-	F1	P+	R+	P-	R-	F1
$\mathcal{D}^{train}$	64.4	89.8	80.5	45.8	66.7	73.6	97.9	91.7	39.1	69.4	46.5	88.8	97.2	79.0	**74.1**
$\mathcal{D}^{train}{}_{SD-rnd}$	64.8	88.1	78.6	47.7	67.0	77.2	97.5	92.0	50.0	**75.5**	45.4	89.4	97.3	78.0	73.4
$\mathcal{D}^{train}{}_{SD-clr}$	66.0	86.8	78.0	51.3	**68.4**	73.4	98.4	93.3	38.3	69.2	43.8	91.8	97.8	75.9	72.4
$\mathcal{D}^{train}{}_{SD-ctg}$	65.5	88.5	79.7	49.2	68.1	74.4	98.1	92.7	41.5	71.0	44.8	91.8	97.9	76.9	73.2
$\mathcal{D}^{train}{}_{SD-gdr}$	64.1	88.5	78.5	45.8	66.1	75.3	96.8	89.2	45.1	72.3	43.9	90.6	97.5	76.3	72.4
$\mathcal{D}^{train}{}_{SD-all}$	65.7	90.8	82.8	48.3	**68.6**	75.8	97.8	92.3	45.9	**73.4**	46.0	92.4	98.0	77.8	**74.1**

Table 2: Precision, Recall and F1-score regarding the use of skill-related words in the captions generated by the BLIP models trained using different TIDA techniques on the different test sets. The two best F1 scores are highlighted in bold.

Skill	Color	Counting	Gender
$\mathcal{D}^{train}$	72.0	88.2	84.1
$\mathcal{D}^{train}{}_{SD-rnd}$	73.0	88.3	84.3
$\mathcal{D}^{train}{}_{SD-clr}$	72.9	88.6	84.7
$\mathcal{D}^{train}{}_{SD-ctg}$	71.6	88.7	84.1
$\mathcal{D}^{train}{}_{SD-gdr}$	71.7	89.0	84.0
$\mathcal{D}^{train}{}_{SD-all}$	71.8	87.7	84.3

Table 3: F1-score for skill probing using the models learned with different targeted data-augmentations

the skill-related information in the image encoding. However, the models are improving in terms of general Image Captioning performances (Table 1), and we saw previously that they are using more frequently targeted words when they should use them (Table 2). We can conclude that TIDA-related improvements are caused by changes in the text decoder rather than the image decoder.

5 Conclusion and Future Work

This paper assesses the effectiveness of generative data augmentation with current diffusion models for improving specific skills of image captioning models. We use the Flickr30k image captioning dataset and ran experiments with BLIP, a recent vision-language state-of-the-art model. We show that TIDA, our targeted image data-augmentation techniques allows for gains regarding classical metrics that are recognized by the community like BLEU or RefCLIPScore. On top of that, we also propose a fine-grained analysis to analyze the results of the model by going beyond the classical opaque metrics by investigating the occurrences of specific semantic words related to the target skills. We found out that TIDA helps the image captioning model to use those words more efficiently. Finally, we investigate the visual part, we probe the representations from the visual encoder and reveal that they do not contain more information on the skill, meaning the improvement relies on the textual decoder.

Our results open several avenues for further research. For instance, it remains unclear why we observe the boost in results on a specific skill when using data-augmentation on another skill. It would also be useful to investigate more in details the reasons of the improvement of performances the text decoder or the visual encoder, or to use a more precised metric powered by a LLM like (Hu et al., 2023).

It would also be useful to investigate more in details the reasons of the gain of performances of the text decoder or the visual encoder, or to using complex interpretable metrics from LLM like the Text-to-Image Faithfulness Evaluation with Question Answering (Hu et al., 2023). It would be to see improvements with text-to-image models known to be better at generating images related to color, counting, like Attend-and-Excite (Chefer et al., 2023) with newer versions of stable diffusion. Finally, we would like to extend our method to Visual Question Answering. Using symbolic knowledge to extract the objects of the image-caption and the relation as implemented in (Riquelme et al., 2020), we can adapt the model to new situations and help to debias a VQA model. Finally, given the recent results of (Azizi et al., 2023), we should run a random data-augmentation on the train set and see whether this procedure may help to improve the results compared with TIDA.

6 Limitations

The focus of this work has been on abstract skills shown to be learned by humans at an early age, but it is not clear which skills are the most important to image captioning in particular or another particular task in general. And it is an empirical study to

determine which skills result in the most improvement in a task. Making it not straightforward to add new skills, requiring thoughtfulness and empirical validation.

In terms of computational cost, TIDA's necessity to generate a number of new examples comparable to the original dataset size using costly neural image generation models signifies it is a challenge to apply to larger datasets and that the technique doesn't scale well to dataset size. And although each generated example can be leveraged many times, the process is heavily limited by the computation capabilities.

Acknowledgments

This work was funded by National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016a. SPICE: Semantic propositional image caption evaluation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9909 LNCS:382–398.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016b. 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.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566.

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2022. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. 2023. Synthetic Data from Diffusion Models Improves ImageNet Classification.

Valentin Barriere. 2017. Hybrid Models for Opinion Analysis in Speech Interactions. In *ICMI*, pages 647–651.

Valentin Barriere, Slim Essid, and Chloé Clavel. 2022. Opinions in Interactions : New Annotations of the SEMAINE Database. *Proceedings of the Language Resources and Evaluation Conference*, (June):7049–7055.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. 2020. Learning Invariances in Neural Networks from Training Data. *Advances in Neural Information Processing Systems (NeurIPS)*, (4):17605–17616.

Marc H Bornstein and Martha E Arterberry. 2003. Recognition, discrimination and categorization of smiling by 5-month-old infants. *Developmental Science*, 6(5):585–599.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2023. Hierarchical organization of language predictions in the brain.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. (i).

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431.

Tsz-Him Cheung and Dit-Yan Yeung. 2022. Adaaug: Learning class-and instance-adaptive data augmentation policies. In *International Conference on Learning Representations*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing Vision-Language Models via Biased Prompts.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13.

Katherine M Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B Tenenbaum. 2022. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *CogSci*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Marie Dacke and Mandyam V Srinivasan. 2008. Evidence for counting in insects. *Animal cognition*, 11(4):683–689.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge Neurons in Pretrained Transformers. 1:8493–8502.

Hank Davis and John Memmott. 1982. Counting behavior in animals: A critical evaluation. *Psychological Bulletin*, 92(3):547.

Felipe del Rio, Julio Hurtado, Cristian Buc, Alvaro Soto, and Vincenzo Lomonaco. 2023. Studying generalization on memory-based methods in continual learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in cognitive sciences*, 8(7):307–314.

Simone Hantke, Nicholas Cummins, and Bjorn Schuller. 2018. What is my dog trying to tell me? the automatic recognition of the context and perceived emotion of dog barks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5134–5138. IEEE.

Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. 2020. Faster autoaugment: Learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 1–16. Springer.

Zexue He, Marco Tulio Ribeiro, and Fereshte Khani. 2023. Targeted Data Generation: Finding and Fixing Model Weaknesses.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. *Proceedings of the European Conference on Computer Vision (ECCV),*, pages 771–787.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. pages 1–19.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, (2014):7514–7528.

Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel. 2019. Population based augmentation: Efficient learning of augmentation policy schedules. In *International Conference on Machine Learning*, pages 2731–2741. PMLR.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. NeuroCounterfactuals: Beyond Minimal-Edit Counterfactuals for Richer Data Augmentation.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An Easier Data Augmentation Technique for Text Classification. In *Findings of ACL: EMNLP 2021*, pages 2748–2754.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 3128–3137.

Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent Human Evaluation for Image Captioning. In *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 3464–3478.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. (2).

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks.

Shari Liu, Bill Pepe, Manasa Ganesh Kumar, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. 2022a. Dangerous ground: One-year-old infants are sensitive to peril in other agents' action plans. *Open Mind*, 6:211–231.

Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, and Andrew Gordon Wilson. 2022b. Learning multimodal data augmentation in feature space. *arXiv preprint arXiv:2212.14453*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and Editing Factual Associations in GPT. (NeurIPS).

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-Editing Memory in a Transformer. (c):1–18.

Ning Miao, Tom Rainforth, Emile Mathieu, Yann Dubois, Yee Whye Teh, Adam Foster, and Hyunjik Kim. 2022. Instance-Specific Augmentation: Capturing Local Invariances. pages 1–15.

Ruslan Mitkov. 2022. *The Oxford handbook of computational linguistics*. Oxford University Press.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models.

Shikhar Murty, Christopher D. Manning, Scott Lundberg, and Marco Tulio Ribeiro. 2022. Fixing Model Bugs with Natural Language Patches. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 11600–11613.

Alexander Quinn 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. In *International Conference on Machine Learning*, pages 16784–16804. PMLR.

OpenAI. 2023. GPT-4 Technical Report. 4:1–100.

Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Cross-modal Attention Congruence Regularization for Vision-Language Relation Alignment.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002. BLEU : a Method for Automatic Evaluation of Machine Translation. (July):311–318.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to Generate Reviews and Discovering Sentiment.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Felipe Riquelme, Alfredo De Goyeneche, Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2020. Explaining VQA predictions using visual grounding and a knowledge base. *Image and Vision Computing*, 101:103968.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. 2020. A benchmark for systematic generalization in grounded language understanding. *Advances in neural information processing systems*, 33:19861–19872.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. 2023. Is a Caption Worth a Thousand Images? A Controlled Study for Representation Learning. In *ICLR*.

Andrew M Saxe, James L McClelland, and Surya Ganguli. 2019. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. 2020. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. 1992. Origins of knowledge. *Psychological review*, 99(4):605.

Elizabeth S Spelke and Katherine D Kinzler. 2007. Core knowledge. *Developmental science*, 10(1):89–96.

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.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings.

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022a. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022b. GIT: A Generative Image-to-text Transformer for Vision and Language. 2:1–49.

Yishi Wang, Karl Ricanek, Cuixian Chen, and Yaw Chang. 2010. Gender classification from infants to seniors. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE.

Jason Wei, Chengyu Huang, Shiqi Xu, and Soroush Vosoughi. 2021. Text Augmentation in a Multi-Task View. In *EACL*, pages 2888–2894.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Olivia Wiles, Isabela Albuquerque, and Sven Gowal. 2022. Discovering Bugs in Vision Models using Off-the-shelf Image Generation and Captioning. 2:1–18.

Fei Xu and Elizabeth S Spelke. 2000. Large number discrimination in 6-month-old infants. *Cognition*, 74(1):B1–B11.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting Image Captioning with Attributes. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:4904–4912.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop*.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

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.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*.

A List of skill-related words

Color We used seven colors: blue, brown, green, grey, orange, pink, purple, red, and yellow. We inverted them randomly.

Counting We used all the numbers from one to six. All the captions only contained written numbers.

Gender For male, we used the words boy, boys, man, men, guy, and guys. They were changed with the words girl, girls, woman, and women.

B Other Image Generators

We generate the images with different techniques. In-Painting mode, in order to change the images the less possible, and another image generator algorithm called Attend-and-Excite (Chefer et al., 2023), in order to stress specific tokens of the sentence used to generated, related to the attribute we want to enhance. Results are in Tables 4 and 5.

B.1 In Painting Model

We ran more experiments with another configuration for image generation that we call Inpainting (INP). It consists of changing only a subpart of the initial image in order to perturbate it. For this configuration, we first segmented the desired object in the scene by using a pretrained ClipSeg model (Lüddecke and Ecker, 2022), by prompting the nominal group of the skill-related word. The segmentation mask was obtained by setting an element-wise threshold of 0.1 in the final output of the model, after applying sigmoid and a min-max normalization. The mask was then dilated using a square kernel of 10 x 10 pixels. The original image was finally inpainted using the pretrained model of (Rombach et al., 2022a).

B.2 Attend-and-Excite

We tried to change the classical stable diffusion by another version called Attend-and-Excite (*AAE*; Chefer et al., 2023), which enhance the classical stable diffusion model to make it better at generating specific attribute.

We used the model described in (Chefer et al., 2023), using as backbone the version 1.5[6] of stable diffusion, with the official implementation of the authors which is also built on top of the Diffusers library. The default parameters were used as default, expect regarding the number maximum of refinement steps, which has been downgraded from 20 to 5.

C Other metrics

Results using other metrics are shown in the section. Table 6 and Table 7 contain respectively the results with Spice and Cider.

D Probing

More results on the probing experiments are shown in Table 8.

[6] https://huggingface.co/runwayml/stable-diffusion-v1-5

Test Train	BLEU@1-4 $\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$	RefCLIPScore $\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$	Spice F1_{clr}	F1_{ctg}	F1_{all}
$\mathcal{D}^{train}$	51.8	44.0	49.9	49.7	79.9	79.3	79.8	80.3	24.1	19.7	20.7
$\mathcal{D}^{train}{}_{INP-clr}$	51.4	44.8	*49.8*	*50.1*	79.8	79.1	79.6	80.1	23.1	20.1	20.4
$\mathcal{D}^{train}{}_{INP-ctg}$	***52.2***	***45.1***	49.3	49.8	80.2	79.3	79.7	80.2	***25.2***	21.3	20.6
$\mathcal{D}^{train}{}_{INP-gdr}$	50.9	42.8	48.3	48.7	***80.3***	*79.6*	***80.2***	***80.5***	23.1	*22.4*	*20.7*
$\mathcal{D}^{train}{}_{INP-all}$	51.3	44.0	49.2	49.5	79.7	79.0	79.6	80.1	23.9	21.3	20.4
$\mathcal{D}^{train}{}_{AAE-clr}$	51.7	42.8	48.7	49.1	80.0	79.0	79.7	80.2	22.6	20.8	20.5
$\mathcal{D}^{train}{}_{AAE-ctg}$	52.1	44.6	49.7	49.9	79.8	79.2	79.7	80.2	24.6	20.3	20.5
$\mathcal{D}^{train}{}_{AAE-gdr}$	51.4	43.5	49.3	49.4	80.1	79.4	80.1	**80.5**	23.7	19.2	20.5
$\mathcal{D}^{train}{}_{AAE-all}$	51.1	43.4	48.8	49.1	79.9	79.5	80.1	80.4	22.9	20.7	**21.0**

Table 4: Average of the BLEU@1-4 scores of the different TIDA-enhanced models on the different test sets. The TIDA models depicted used different image generation strategies: *SD* uses Stable Diffusion, *AAE* Attend-and-Excite, and *INP* Inpaiting. The first line contains the performance of the model trained with the Vanilla train set. Then, the first to third line of each TIDA model contain the results of the model trained with data-augmentation on the color, counting, and gender skills, respectively. And, the last line of each, depicts the results of the model trained with all three types of data-augmentation. The scores in bold are the best scores on each test set, while the scores in italic are the best scores of each of the models trained with (skill-related) data-augmentation.

Skill Train	Color P+	R+	P-	R-	F1	Counting P+	R+	P-	R-	F1	Gender P+	R+	P-	R-	F1
$\mathcal{D}^{train}$	64.4	89.8	80.5	45.8	66.7	73.6	97.9	91.7	39.1	69.4	46.5	88.8	97.2	79.0	74.1
$\mathcal{D}^{train}{}_{INP-clr}$	63.6	91.2	81.7	42.9	65.6	73.3	98.4	93.3	38.0	69.0	45.1	89.4	97.3	77.7	73.2
$\mathcal{D}^{train}{}_{INP-ctg}$	64.7	87.9	78.4	47.7	66.9	74.5	96.8	88.6	42.6	70.9	42.6	91.8	97.8	74.7	71.5
$\mathcal{D}^{train}{}_{INP-gdr}$	63.1	88.7	77.8	43.3	64.7	74.4	96.8	88.6	42.3	70.7	44.7	90.0	97.4	77.2	73.0
$\mathcal{D}^{train}{}_{INP-all}$	64.5	88.9	79.4	46.7	66.8	74.3	97.8	91.6	41.5	70.8	45.8	92.9	98.2	77.5	74.0
$\mathcal{D}^{train}{}_{AAE-clr}$	62.8	90.4	79.9	41.6	64.5	74.3	97.5	90.5	41.5	70.6	47.4	91.2	97.8	79.3	75.0
$\mathcal{D}^{train}{}_{AAE-ctg}$	64.0	88.7	78.6	45.4	65.9	74.0	98.4	93.6	40.2	70.4	47.3	91.2	97.8	79.2	74.9
$\mathcal{D}^{train}{}_{AAE-gdr}$	63.9	90.0	80.3	44.4	65.9	74.3	97.8	91.6	41.5	70.8	42.9	90.0	97.4	75.4	71.5
$\mathcal{D}^{train}{}_{AAE-all}$	64.4	90.6	81.5	45.2	66.7	75.4	97.3	90.7	45.1	72.6	48.6	90.6	97.7	80.4	75.7

Table 5: Precision, Recall and F1-score regarding the use of skill-related words in the captions generated by the BLIP models trained using different TIDA techniques on the different test sets

Test Train	$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$
$\mathcal{D}^{train}$	21.3	18.5	20.3	20.7
$\mathcal{D}^{train}{}_{SD-rnd}$	21.4	18.2	20.1	20.6
$\mathcal{D}^{train}{}_{SD-clr}$	20.9	17.9	19.7	20.2
$\mathcal{D}^{train}{}_{SD-ctg}$	21.0	18.2	20.0	20.4
$\mathcal{D}^{train}{}_{SD-gdr}$	20.8	18.8	19.9	20.6
$\mathcal{D}^{train}{}_{SD-all}$	21.0	19.3	20.3	21.0
$\mathcal{D}^{train}{}_{AAE-clr}$	20.8	18.0	19.8	20.5
$\mathcal{D}^{train}{}_{AAE-ctg}$	21.1	18.6	20.0	20.5
$\mathcal{D}^{train}{}_{AAE-gdr}$	21.0	18.3	19.9	20.5
$\mathcal{D}^{train}{}_{AAE-all}$	21.2	18.7	20.3	21.0
$\mathcal{D}^{train}{}_{INP-clr}$	20.7	18.4	19.9	20.4
$\mathcal{D}^{train}{}_{INP-ctg}$	21.6	18.8	20.2	20.6
$\mathcal{D}^{train}{}_{INP-gdr}$	21.1	18.9	20.1	20.7
$\mathcal{D}^{train}{}_{INP-all}$	20.9	18.4	19.9	20.4

Table 6: Average of the Spice F1 scores of the different models on the different test sets

Test Train	$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$
$\mathcal{D}^{train}$	102.5	81.1	95.3	99.6
$\mathcal{D}^{train}{}_{SD-rnd}$	100.9	81.7	94.9	99.3
$\mathcal{D}^{train}{}_{SD-clr}$	102.2	80.3	94.0	98.8
$\mathcal{D}^{train}{}_{SD-ctg}$	102.2	82.3	93.9	99.0
$\mathcal{D}^{train}{}_{SD-gdr}$	100.1	81.9	92.7	98.0
$\mathcal{D}^{train}{}_{SD-all}$	101.0	81.4	95.7	101.5
$\mathcal{D}^{train}{}_{AAE-clr}$	102.2	77.8	92.7	98.0
$\mathcal{D}^{train}{}_{AAE-ctg}$	101.7	82.0	95.1	100.5
$\mathcal{D}^{train}{}_{AAE-gdr}$	99.5	78.1	93.6	98.0
$\mathcal{D}^{train}{}_{AAE-all}$	99.5	78.5	92.8	98.2
$\mathcal{D}^{train}{}_{INP-clr}$	100.8	82.9	95.3	100.5
$\mathcal{D}^{train}{}_{INP-ctg}$	104.5	83.7	94.7	99.8
$\mathcal{D}^{train}{}_{INP-gdr}$	101.7	80.6	94.1	99.0
$\mathcal{D}^{train}{}_{INP-all}$	100.7	82.3	94.5	99.4

Table 7: Average of the Cider scores of the different models on the different test sets

Skill	**Color**			**Counting**			**Gender**		
Train	P	R	F1	P	R	F1	P	R	F1
$\mathcal{D}^{train}$	67.5	77.2	72.0	87.9	88.6	88.2	83.1	85.1	84.1
$\mathcal{D}^{train}{}_{SD-rnd}$	70.7	75.4	73.0	86.1	90.5	88.3	83.2	85.4	84.3
$\mathcal{D}^{train}{}_{SD-clr}$	69.1	77.2	72.9	86.0	91.4	88.6	83.3	86.2	84.7
$\mathcal{D}^{train}{}_{SD-ctg}$	66.3	77.8	71.6	85.1	92.6	88.7	82.6	85.7	84.1
$\mathcal{D}^{train}{}_{SD-gdr}$	67.8	76.1	71.7	85.5	92.7	89.0	83.9	84.2	84.0
$\mathcal{D}^{train}{}_{SD-all}$	60.1	89.1	71.8	86.8	88.6	87.7	83.3	85.3	84.3
$\mathcal{D}^{train}{}_{AAE-clr}$	68.5	75.9	72.0	86.7	89.2	88.0	84.1	86.5	85.3
$\mathcal{D}^{train}{}_{AAE-ctg}$	65.3	83.5	73.3	86.1	90.6	88.3	82.9	86.7	84.7
$\mathcal{D}^{train}{}_{AAE-gdr}$	71.8	73.7	72.7	85.2	91.9	88.4	84.0	86.7	85.3
$\mathcal{D}^{train}{}_{AAE-all}$	72.5	75.6	74.0	89.0	90.2	89.6	81.4	87.8	84.5
$\mathcal{D}^{train}{}_{INP-clr}$	63.7	80.5	71.1	84.3	91.0	87.5	84.6	83.4	84.0
$\mathcal{D}^{train}{}_{INP-ctg}$	67.6	79.1	72.9	88.1	89.1	88.6	83.9	84.8	84.3
$\mathcal{D}^{train}{}_{INP-gdr}$	66.0	81.0	72.7	88.6	89.5	89.0	82.2	85.9	84.0
$\mathcal{D}^{train}{}_{INP-all}$	66.4	79.4	72.3	87.4	91.1	89.2	85.9	83.7	84.8

Table 8: Skill Probing