Guiding image captioning models toward more specific captions

Simon Kornblith[1] Lala Li[1] Zirui Wang[2*] Thao Nguyen[3†]

[1]Google DeepMind [2]Apple AI/ML [3]University of Washington

Abstract

Image captioning is conventionally formulated as the task of generating captions for images that match the distribution of reference image-caption pairs. However, reference captions in standard captioning datasets are short and may not uniquely identify the images they describe. These problems are further exacerbated when models are trained directly on image-alt text pairs collected from the internet. In this work, we show that it is possible to generate more specific captions with minimal changes to the training process. We implement classifier-free guidance [14] for an autoregressive captioning model by fine-tuning it to estimate both conditional and unconditional distributions over captions. The guidance scale applied at decoding controls a trade-off between maximizing $p(\text{caption}|\text{image})$ and $p(\text{image}|\text{caption})$. Compared to standard greedy decoding, decoding with a guidance scale of 2 substantially improves reference-free metrics such as CLIPScore (0.808 vs. 0.775) and caption→image retrieval performance in the CLIP embedding space (recall@1 44.6% vs. 26.5%), but worsens standard reference-based captioning metrics (e.g., CIDEr 78.6 vs 126.1). We further explore the use of language models to guide the decoding process, obtaining small improvements over the Pareto frontier of reference-free vs. reference-based captioning metrics that arises from classifier-free guidance, and substantially improving the quality of captions generated from a model trained only on minimally curated web data.

1. Introduction

Image captioning is both a difficult task for computer vision systems to perform and a difficult task to evaluate. Although automated captioning metrics rank the best captioning systems higher than humans, human raters still show a strong preference for human-generated captions [20], suggesting shortcomings in both captioning models and metrics. One shortcoming relates to the lack of specificity in generated captions. Conventional maximum likelihood-

γ=1.0 a man riding a blue motorcycle on a dirt road
γ=1.5 a man riding a blue motorcycle on a straw bale
γ=2.0 rider on blue suzuki motorcycle near straw bales
γ=3.0 racer on blue suzuki motorcycle leaning up against straw bales
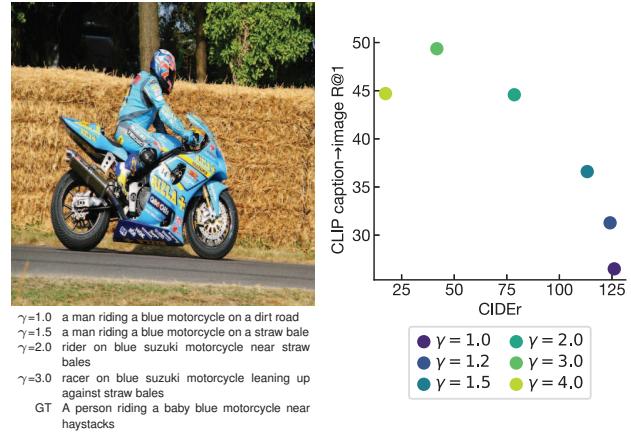GT A person riding a baby blue motorcycle near haystacks

Figure 1. Using classifier-free guidance ($\gamma > 1$) results in more specific captions that are farther from the reference distribution. Left: Example of captions generated at different guidance scales for a single image. Right: Caption→image recall@1 with CLIP ViT-B/32 vs. CIDEr score, for captions generated with different guidance scales γ on MS-COCO. Higher scales improve retrieval accuracy at the expense of CIDEr.

based image captioning models attempt to generate captions such that the $p(\text{caption}|\text{image})$ is high. However, captions from the ground truth distribution are often non-specific, e.g., human annotators will usually describe a German Shepard only as a dog. Moreover, previous work has emphasized "reference-based" captioning metrics that measure the match between generated captions and human-provided ground truth captions [28, 23, 42]. These metrics intrinsically penalize captions that are more specific than ground truth.

In this work, we explore strategies to guide image captioning models to produce more specific captions by modifying the decoding distribution, and explore the trade-offs in captioning metrics that result. We first investigate the application of classifier-free guidance (CFG) [14] to image captioning with autoregressive models. Classifier-free guidance increases $p(\text{image}|\text{caption})$ at the expense of $p(\text{caption}|\text{image})$. Although CFG hurts reference-based image captioning metrics such as BLEU [28], ROUGE [23], and CIDEr [42], it improves "reference-free" metrics that measure captions' specificity via the similarity between the

image and the generated caption in the embedding space of image-text models [13] or caption→image retrieval performance. Qualitatively, we find that captions generated with CFG are more specific than both the ground truth captions and captions generated without CFG, but they are less grammatical, particularly at high CFG scales.

Beyond classifier-free guidance, we experiment with guiding image captioning models using the probability distribution obtained from a few shot-prompted language model (LM). We find that using a language model to guide a captioning model trained on MS-COCO [24] with descriptive manually written captions can allow it to achieve slightly better trade-offs between reference-free vs. reference-based captioning metrics than those observed with CFG. LM guidance also substantially improves the captions produced by a model trained exclusively on minimally curated web data. Although this model achieves a CIDEr score of only 21.8 without guidance, this CIDEr score improves to 57.4 when guided by a language model prompted with 20 captions from the MS-COCO training set.

In summary, our contributions are as follows:

- We propose two strategies to guide image captioning models to produce more specific captions: classifier-free guidance and language model guidance.

- We demonstrate that classifier-free guidance yields captions that are closer to the corresponding image in the embedding space of image-text models, but are farther from human-provided reference captions.

- We show that language model guidance can alter caption styles, substantially improving captions produced by a model trained only on minimal curated web data and marginally improving the trade-off between captioning metrics observed with classifier-free guidance.

2. Related work

Measuring specificity of captions. Early work using neural networks for image captioning found that models have a propensity to regurgitate captions from their training data, and as a result, the generated captions are not descriptive enough to uniquely identify images [43, 11]. To address this shortcoming, Lindh et al. [25] proposed to use caption→image recall with an image retrieval model to examine whether images can be retrieved from generated captions, and further attempt to differentiate through this retrieval process to train a captioning model. Their approach marginally improves retrieval accuracy, but worsens reference-based captioning metrics. More recent work has adopted approaches to evaluate the specificity of captions based on the CLIP image-text model [30]. Hessel et al. [13] propose CLIPScore, an image captioning metric based on the cosine similarity between CLIP embeddings of the image and the generated caption. Kasai et al. [20] re-

port that CLIPScore-based metrics align better with human judgments compared to reference-based captioning metrics.

Improving specificity of captions. Recent work has attempted to train caption models to directly maximize the similarity of captions with corresponding images in the CLIP embedding space using reinforcement learning. This work has found that adding such losses worsens standard reference-based captioning metrics but improve similarity and retrieval in the CLIP embedding space [16, 6, 51], similar to our observations regarding CFG. Wen et al. [44] attempt to generate prompts for text-to-image generative models that correspond to specific images without a captioning model, by directly optimizing the similarity between the text and image in the CLIP embedding space using a gradient-based discrete optimization procedure, but the resulting text is not grammatical.

Other work has attempted to generate more descriptive captions through different means. Dense captioning [46] aims to detect and caption all objects in an image, but concatenating all of these captions leads to long and unnatural captions, whereas CFG produces single-sentence captions. The Localized Narratives dataset [29] contains visually grounded captions for MS-COCO images collected through voice annotation. These captions are substantially more descriptive than the captions in the MS-COCO dataset and can be used for model training. Two recent papers propose to combine multiple captions from a single model [5] or outputs of different vision models [34], and then combine them using a language model. The resulting captions are longer, but achieve greater caption→image recall.

Captioning from uncurated data. In Section 4.2, we explore the use of LM guidance for captioning with access to uncurated image-text data from the web and a small number of captions but not images from the target distribution. This setting, which does not rely on aligned images and captions from the target distribution, is often referred to as "zero-shot" captioning, and previous work has pursued a number of alternative approaches. Flamingo [3] and CM3 [1] perform zero-shot captioning by pretraining on interleaved image/text data. MAGIC [39] and ZeroCap [41] generate captions using a combination of guidance from CLIP and a large language model. Other recent work adapts CLIP to perform captioning by training a text decoder using only captions, with no corresponding images [27, 22].

Classifier-free guidance. CFG is widely used in diffusion-based and autoregressive text-to-image models [26, 32, 35, 33, 12, 48]. Because of the popularity of the combination of CFG and diffusion, previous work that has performed image captioning with diffusion models has also examined the use of CFG. This work finds either no benefit to using CFG [45] or a small and inconsistent benefit that appears to vary with minor changes in training settings [52]. However, these studies do not seek to generate

more specific captions, and thus measure only reference-based captioning metrics, which we likewise find do not benefit from CFG. Concurrently with out work, [36] propose to use classifier-free guidance to improve prompt following in large language models.

3. Methods

3.1. Classifier-free guidance for image captioning

Let x be an image caption and y be the corresponding image. A standard captioning model aims to model the likelihood $p(\mathrm{x}|\mathrm{y})$, factorized autoregressively in terms of the probability of each token given previous tokens

$$p(\mathrm{x}|\mathrm{y}) = p(\mathrm{x}_n|\mathrm{x}_{n-1}, \dots \mathrm{x}_1, \mathrm{y}) \dots p(\mathrm{x}_1|\mathrm{y}). \quad (1)$$

The network is trained so that its output distribution $q_{\boldsymbol{\theta}}(\mathrm{x}_n|\mathrm{x}_{n-1}, \dots \mathrm{x}_1, \mathrm{y}) \overset{\text{def}}{=} \text{softmax}(f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \dots \mathrm{x}_1, \mathrm{y}))$ approximates $p(\mathrm{x}_n|\mathrm{x}_{n-1}, \dots \mathrm{x}_1, \mathrm{y})$. At inference time, one typically uses beam search or greedy decoding to produce a caption that has a particularly high probability. In this work, we use greedy decoding because it is the more common choice and it is also simpler to implement.

Classifier-free guidance (CFG) [14] modifies the inference procedure so it instead aims to achieve high values of

$$l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y}) \overset{\text{def}}{=} p(\mathrm{x}) \left(\frac{p(\mathrm{x}|\mathrm{y})}{p(\mathrm{x})} \right)^{\gamma} \propto p(\mathrm{x})p(\mathrm{y}|\mathrm{x})^{\gamma}, \quad (2)$$

where proportionality holds because $p(\mathrm{x}|\mathrm{y})/p(\mathrm{x}) = p(\mathrm{y}|\mathrm{x})/p(\mathrm{y})$ and $p(\mathrm{y})$ is fixed. The parameter γ is called the guidance scale and controls the trade-off between maximization of $p(\mathrm{x}|\mathrm{y})$ and $p(\mathrm{y}|\mathrm{x})$. When $\gamma = 1$, $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y}) = p(\mathrm{x}|\mathrm{y})$ and guidance has no effect. Setting $\gamma > 1$ inflates the probability of the image given the caption $p(\mathrm{y}|\mathrm{x})$ relative to the unconditional probability of the caption $p(\mathrm{x})$.

Ho and Salimans [14] originally proposed CFG in the context of diffusion models, which estimate the score functions $\nabla \log p(\mathrm{x}|\mathrm{y})$ and $\nabla \log p(\mathrm{x})$. Although $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y})$ factorizes autoregressively, it is not a normalized probability distribution, so it is not entirely clear how one should sample tokens when performing autoregressive generation. Crowson [8] suggested to sample from

$$\tilde{q}_{\boldsymbol{\theta},\gamma}(\mathrm{x}_n|\mathrm{x}_{n-1}, \dots, \mathrm{x}_1, \mathrm{y}) \overset{\text{def}}{=} \text{softmax}(f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \dots, \mathrm{x}_1, \mathbf{0})$$
$$+ \gamma(f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \dots, \mathrm{x}_1, \mathrm{y}) - f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \dots, \mathrm{x}_1, \mathbf{0}))), \quad (3)$$

where $f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \dots, \mathrm{x}_1, \mathbf{0})$ are logits generated by the model without conditioning, usually by passing zeros in place of the conditioning information. This formulation has been successfully applied in autoregressive image models [12, 48]. In our experiments, since we decode greedily, i.e., at each step we take the token that maximizes $\tilde{q}_{\boldsymbol{\theta},\gamma}(\mathrm{x}_n|\mathrm{x}_{n-1}, \dots, \mathrm{x}_1, \mathrm{y})$ and thus $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y})$, any form of normalization of the per-step sampling distribution would produce the same captions. We provide pseudocode in Appendix A.1. Crucially, the only change required to train-

ing is to learn an unconditional model of captions alongside the image-conditonal model, which we accomplish by randomly masking the image embeddings.

3.2. Language model guidance

Inspired by classifier-free guidance, we consider *language model (LM) guidance*, which attempts to maximize

$$l'_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y}) \overset{\text{def}}{=} q(\mathrm{x}) \left(\frac{p(\mathrm{x}|\mathrm{y})^{\alpha}}{p(\mathrm{x})^{\beta}} \right), \quad (4)$$

where $p(x)$ and $p(x|y)$ are obtained from a captioning model as in CFG but $q(\mathrm{x})$ is obtained from a language model that was trained independently (but with the same vocabulary) on a large text corpus. The quantity $p(\mathrm{x}|\mathrm{y})/p(\mathrm{x}) = p(\mathrm{x}, \mathrm{y})/(p(\mathrm{x})p(\mathrm{y}))$ measures the strength of the association between a caption and an image; its logarithm is the pointwise mutual information (PMI). LM guidance relies on the assumption that, even for large shifts in the prior distribution of captions $p(\mathrm{x})$, the shift in PMI will be small. Empirically, we obtain better results by allowing different exponents for the numerator and denominator, with $\alpha > \beta$. This decoupling resembles PMI^k [9], which reduces the bias of PMI toward rare associations. We provide a more detailed derivation in Appendix A.2.

We investigate two applications of LM guidance. First, we combine a captioning model fine-tuned on MS-COCO with a LM prompted with manually written descriptive captions to alter the style of the captions the model produces. The manually written prompts are shown in Appendix A.4. Second, we combine a captioning model trained only on low-quality web data with a LM prompted with varying numbers of examples from the MS-COCO training set to evaluate the ability of LM guidance to elicit higher-quality captions without high-quality paired data. We randomly select a different set of captions for each minibatch of four test examples. In both cases, we separate the captions with two newlines. Because this format leads the LM to place probability mass on the newline token to end the caption, we transfer the probability mass from the newline token to the EOS token. See Appendix A.3 for pseudocode.

3.3. Models and training

Our captioning model is a "bottleneck" variant of CoCa-Base [47], which combines a contrastive loss with a captioning loss to simultaneously learn aligned image and text embeddings as well as a captioner. The architecture consists of an image encoder, a unimodal text decoder, and a multimodal text decoder, each of which are Transformers with 12 layers, 768 hidden dimensions, an MLP of size 3072, and 12 self-attention heads, matching BERT$_{\text{BASE}}$ [10] and GPT-1 [31]. The image encoder is a ViT-B/18 that processes 288×288 input and produces an embedding such that images are embedded close to their corresponding text.

CoCa's multimodal text decoder processes the representations of the image encoder to produce a caption. Whereas [47] conditions the multimodal text decoder using cross-attention to pooled representations, our bottleneck variant uses only the contrastive image embedding. Appendix A.5 shows a diagram of the resulting architecture. We adopt this bottleneck variant because of its simplicity and the conceptual appeal: When CFG is used, the captioner's role is to invert the image embedding, providing a caption that, when embedded by the text encoder, lies close to it. However, as we show in Appendix B.1, this choice of the bottleneck model is not critical, and CFG is equally effective with the standard CoCa architecture with attention pooling.

For CFG experiments, we pretrain our model on an image-text dataset comprising images from the JFT-5B dataset [40, 49] paired with their corresponding label names substituted into a randomly selected prompt from the list provided by Radford et al. [30], web images paired with noisy alt text from the ALIGN dataset [17], and a small amount of data from other sources. We do not mask conditioning information during pretraining.[1] We then fine-tune on the combined MS-COCO train and Karpathy validation splits [18] using Adam with batch size 128. We linearly warm up to a learning rate of 1×10^{-5} over the first $1,000$ steps and linearly decay to zero over the rest of training. We vary $\gamma \in \{1.0, 1.2, 1.5, 2.0, 3.0, 4.0\}$, conditioning masking proportion in $\{0.0, 0.25, 0.5, 0.75\}$, and numbers of steps in $\{5,000, 10,000, 20,000, 50,000\}$. We report results from the model trained for $20,000$ steps with masking proportion 0.5, which achieves near-optimal results, in Tables 1 and B.4, and sample example captions from it. To ensure that results generalize across datasets, we also experiment with a model fine-tuned on Conceptual Captions [37] for 100,000 steps with masking proportion 0.5.

For LM guidance experiments, we pretrain on the JFT-5B and ALIGN datasets, again following the recipe of [47]. For zero-shot captioning experiments, we fine-tune this model on the same datasets for an additional 50,000 steps with conditioning masking proportion of 0.5 to improve our ability to sample unconditionally. For LM guidance on MS-COCO, we first fine-tune on ALIGN, JFT-5B images backcaptioned by an MS-COCO fine-tuned CoCa-2B model, and a small amount of internal data before fine-tuning on MS-COCO. Our language model is a variant of Primer [38] with 2 billion parameters, trained on a similar dataset to that used to train PaLM [7].

3.4. Evaluation

We adopt the standard reference-based captioning metrics BLEU-4, METEOR, ROUGE, and CIDEr, as well

as reference-free captioning metrics based on CLIP ViT-B/32 [30]. The first reference-free captioning metric is CLIPScore [13], which is defined as $\texttt{CLIP-S}(c, v) = 2.5 \cdot \max(\cos(c, v), 0)$ where c and v are the CLIP embeddings of the caption and image respectively. The second reference-free metric measures the accuracy with which we can retrieve an image from the generated caption within a given test split by taking the k nearest neighbors of the caption in the CLIP embedding space. Because recall@k for $k > 1$ is highly correlated with recall@1 (R@5: $r = 0.99$, R@10: $r = 0.98$), we plot only recall@1. We additionally report RefOnlyCLIP-S, a reference-based metric that uses the CLIP text encoder to compute the similarity of CLIP embeddings of the generated captions with embeddings of ground truth captions, and RefCLIP-S, which takes the average of the per-image harmonic means of CLIP-S and RefOnlyCLIP-S [13]. Unless otherwise stated, all evaluation is performed on the MS-COCO Karpathy test split [18].

4. Results

4.1. Classifier-free guidance

We first investigate the trade-off between reference-based and reference-free image captioning metrics as a function of guidance scale. Because different guidance scales and metrics could conceivably benefit from different fine-tuning hyperparameter combinations, we plot all results from our hyperparameter grid in Figure 2. Although standard greedy decoding ($\gamma = 1.0$) produces the highest CIDEr, METEOR, ROUGE, and BLEU-4 scores, higher guidance weights consistently yield higher values of reference-free captioning metrics. In particular, $\gamma = 3.0$ offers both the best caption→image recall and the best CLIP-Score.

Table 1 compares our results, obtained from a single model evaluated at different guidance scales, with previous work that reports either CLIPScore or CLIP ViT-B/32 caption→image retrieval performance. Although our model is trained with standard cross-entropy loss rather than a CLIP-based loss and our pretraining dataset is distinct from CLIP's, sampling from our model with CFG yields higher CLIPScores than all other models trained without CLIP-based losses, and better CLIP caption→image retrieval even when compared with models that use CLIP-based losses.

We present examples of captions generated at different CFG scales in Figure 3. Higher CFG strengths lead to more descriptive captions. At $\gamma = 1.0$, the central object in the top left image is described as a "car" as in the ground truth caption, whereas at $\gamma > 1.0$ it is a "station wagon." Similarly, at low CFG strengths, the birds in the center image are described simply as "birds," whereas at $\gamma = 2.0$ they become "crested cranes." However, at $\gamma = 3.0$, captions clearly become less grammatical, containing repeated

---

[1] We find that passing an all-zero image embedding to the pretrained model yields samples that resemble the unconditional distribution, suggesting that it implicitly learns to model the unconditional distribution.
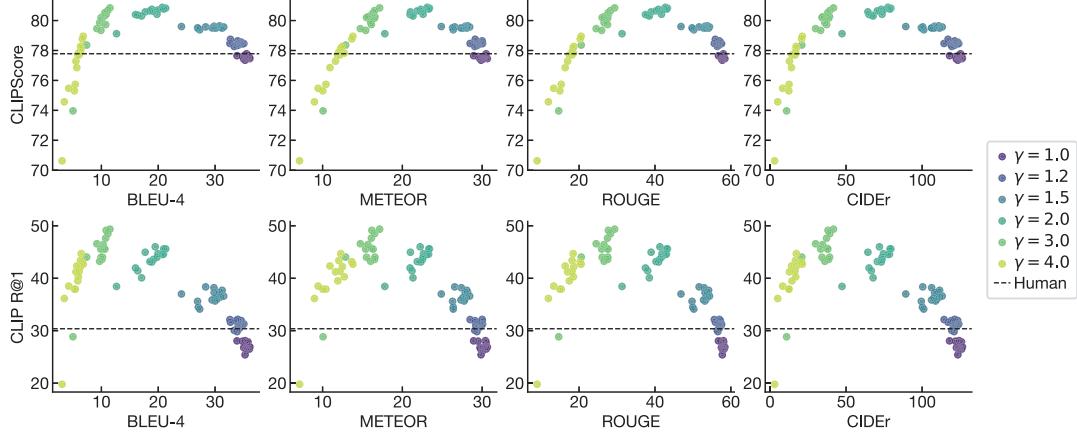
Figure 2. Classifier-free guidance controls a trade off between reference-free and reference-based captioning metrics. Each point reflects a model trained with a different hyperparameter combination; each color represents a γ value used to decode. Models are evaluated with different guidance scales γ, using reference-free captioning metrics based on CLIP ViT-B/32 (y-axes; top: CLIPScore, bottom: recall@1) and reference-based captioning metrics (x-axes). The dashed line reflects the value of the reference-free captioning metric for the ground-truth captions obtained from MS-COCO.

Model	Reference-Based Metrics					Reference-Free Metrics				
	BLEU-4	METEOR	ROUGE	CIDEr	RefOnlyCLIP-S	CLIP-S	R@1	R@5	R@10	RefCLIP-S
Models trained with CLIP features or losses:										
CLIP-Captioner [4]	38.7	29.3	58.6	126.0	0.811	0.754				0.814
UMT-BITG [16]	37.3	28.2	57.9	122.6		0.772				
X-LAN+SCST+GEG [51]	36.5	28.7	57.5	121.7			28.1	50.3	67.2	
CIDEr + CLIP-S Reward [6]	37.7	28.8	58.3	124.6		0.772	24.4	50.2	63.1	
CLIP-S Reward [6]	6.2	18.7	31.6	11.2		0.860	42.5	71.6	82.2	
ZeroCap [41]	2.6	11.5		14.6		0.87				0.79
Models trained without access to CLIP:										
UMT-BITG w/o CLIP loss [16]	37.6	28.3	58.1	122.5		0.725				
VinVL-large [50]	**41.0**	**30.9**	59.4*	**140.9**	**0.91***	0.78*				**0.84***
Ours ($\gamma = 1.0$)	36.1	30.5	58.2	126.1	0.900	0.775	26.5	51.9	64.1	0.830
Ours ($\gamma = 1.2$)	35.1	30.0	57.5	124.1	0.899	0.785	31.3	57.4	69.3	0.835
Ours ($\gamma = 1.5$)	31.5	28.4	54.4	113.2	0.891	0.796	36.6	64.0	75.0	**0.838**
Ours ($\gamma = 2.0$)	20.9	23.3	43.0	78.6	0.862	**0.808**	44.6	71.7	81.7	0.831
Ours ($\gamma = 3.0$)	11.5	17.1	29.4	41.7	0.820	**0.808**	49.4	75.7	**84.7**	0.811
Ours ($\gamma = 4.0$)	6.5	12.3	18.4	17.3	0.766	0.782	44.7	71.3	80.9	0.771

Table 1. Quantitative comparison of our approach with results from previous work that reports CLIP-based metrics. For VinVL-large, * indicates metrics from [19].
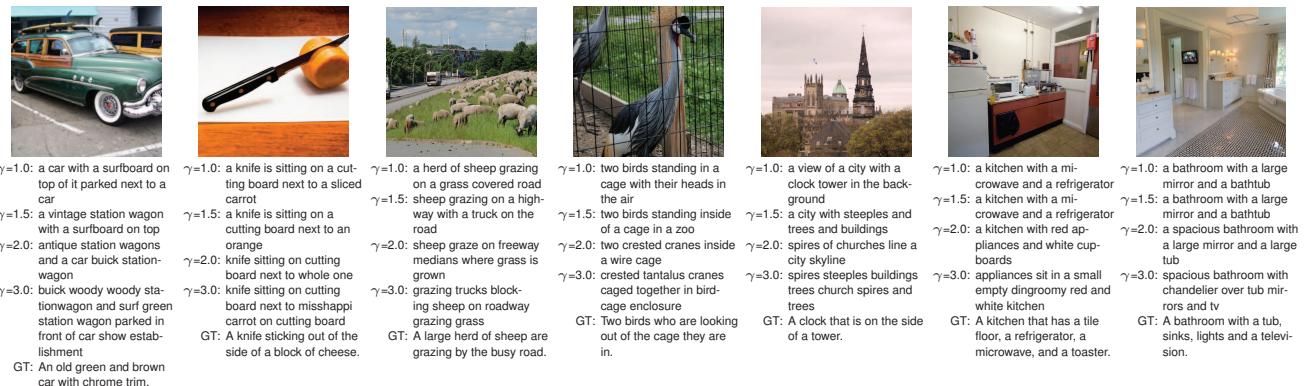

Figure 3. Caption descriptiveness increases with CFG strength, but high CFG strengths produce agrammatical captions. Here we show examples of captions generated with different classifier-free guidance scales, for randomly selected images without human faces from the MS-COCO Karpathy test split. Captions labeled $\gamma = 1.0$ are obtained without CFG; $\gamma > 1$ uses CFG; GT = ground truth.

words ("woody woody") and nonsense words ("misshappi", "dingroomy"). Figure 4 shows captions obtained with and without CFG next to the top 5 closest images in the embedding space of CoCa 2B [47],[2] where it is clear that CFG adds details to captions that help to distinguish them from other captions in the test split. We provide additional examples in Appendix C.

To provide additional quantitative assessments of the specificity of elicited captions, we perform two additional evaluations, described further in Appendix B.2. First, we generate captions for the Stanford Dogs [21] test set, which consists of 8,580 images in total of 120 breeds of dogs, and examine their properties. Without guidance, only 1.9% of captions contain one of the 120 breed names, whereas at $\gamma = 2.0$, 42.4% do. The percentage of these breed names that are correct changes little, from 61.7% without guidance to 58.5% at $\gamma = 2.0$. Second, we performed a human evaluation comparing captions of MS-COCO test set images obtained without guidance and at $\gamma = 2.0$. We asked subjects to select the caption that is "better" and "more descriptive" or to indicate that they are both equal. When we asked these questions separately, we found that the two sets of captions are statistically indistinguishable. However, when asking both questions on the same survey, we found that captions generated without guidance are slightly "better" (50.5% vs. 46.6%, $p = 0.006$, binomial test) but captions generated at $\gamma = 2.0$ are "more descriptive" (52.7% vs. 45.8%, $p = 1 \times 10^{-6}$).

To validate the reliability of our results, we further measure the impact of CFG on three additional datasets, **nocaps** [2], Flickr-8k [15], and Conceptual Captions (CC3M) [37], as well as with alternative retrieval models. **nocaps** is a test set for captioning models with objects not present in MS-COCO; Flickr-8k is a small captioning dataset collected using a different procedure than MS-COCO; and Conceptual Captions is a set of 3.3M captions collected from filtered alt-text. We fine-tune the bottleneck CoCa-Base model directly on CC3M, and use our model fine-tuned on MS-COCO to caption images on **nocaps** and Flickr-8K. As shown in Figure 5, we find trade-offs between reference-based and reference-free captioning metrics similar to those above. In Appendix B.3, we report reference-free captioning metrics on MS-COCO computed with two additional retrieval models: the pretrained CoCa 2B model from [47] and the fine-tuned CoCa Base model that we use to generate captions. With both models, CFG substantially increases recall, in line with results obtained with CLIP ViT-B/32.

Although CFG produces captions that are more successful at uniquely identifying images than decoding from the conditional distribution, caption lengths are similar for

Figure 4. Captions generated with CFG contain specific details that improve retrieval. For each reference image (far left), we show captions at $\gamma = 1.0$ (no guidance) and $\gamma = 2.0$. To the right, we show the closest images to each caption in the CoCa embedding space. Reference images are selected at random subject to the constraints that the closest image differs between γ values and there are no identifiable human faces.
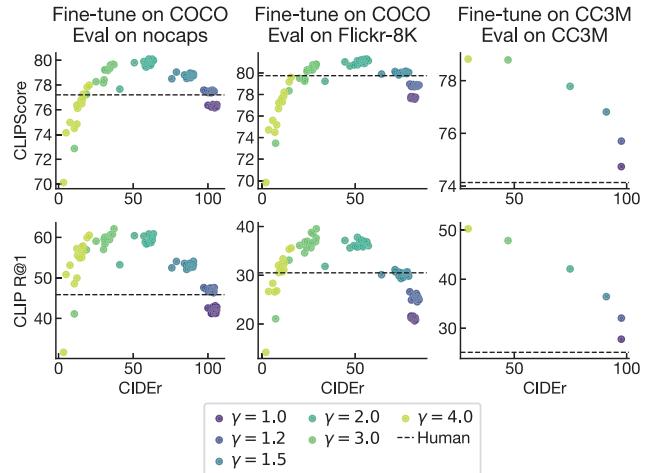

Figure 5. CFG also yields trade-offs between captioning metrics on nocaps, Flickr-8K, and CC3M.

---

[2]We use CoCa 2B rather than CLIP because, quantitatively and qualitatively, it provides better retrieval results both with and without guidance.

γ	Words	Characters
1.0	9.6 ± 1.4	44.2 ± 7.2
1.2	9.6 ± 1.4	44.7 ± 7.4
1.5	9.4 ± 1.4	45.7 ± 7.8
2.0	9.3 ± 2.4	50.3 ± 18.6
3.0	10.7 ± 7.6	69.0 ± 56.1
4.0	19.9 ± 16.9	161.2 ± 140.0

Table 2. Moderate CFG scales do not substantially change caption lengths, although higher CFG scales result in longer captions. Numbers are mean $\pm$ standard deviation.
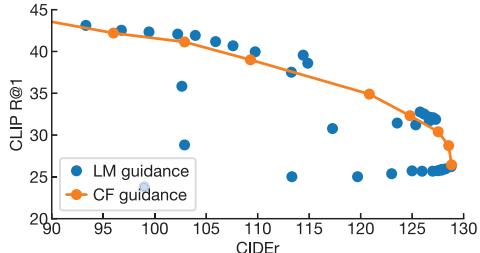

Figure 6. Language model guidance produces captions that slightly exceed the Pareto frontier of CIDEr vs. caption→image retrieval accuracy on MS-COCO.

$\gamma \in [1, 2]$, as shown in Table 2. Thus, at low guidance strengths, CFG improves recall by making more efficient use of words, rather than by producing more verbose captions. Higher CFG strengths lead to longer captions but, as described above, these captions are agrammatical and contain nonsense words.

4.2. Language model guidance

We first experiment with guiding a captioning model fine-tuned on MS-COCO to produce more descriptive captions using a language model prompted with manually written prompts. We first manually wrote a prompt containing 10 descriptive captions of COCO test set images (Appendix A.4). We then sweep over $\alpha \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15\}$ and $\beta \in \{0, \alpha/4, \alpha/2, 3/4\alpha, \alpha\}$, and compare the resulting retrieval/CIDEr trade-off to that produced by the same model with CFG. We observe that it is possible to obtain small improvements upon the Pareto frontier provided by CFG, as shown in Figure 6. With $\alpha = 5$, $\beta = -5/2$, LM guidance achieves CLIP ViT-B/32 R@1 of 39.6% and CIDEr of 114.4, whereas CFG with $\gamma = 1.6$ is worse on both metrics, achieving R@1 of 39.0% and CIDEr of 109.3.

We further experiment with using prompting to control the captioner using a manually written prompt of 25 captions in the form of "a photo of NUMBER OBJECTS" (e.g., "a photo of eight apples"; see Appendix A.4). With $\alpha = \beta = 1$, the guided model is able to match this format and counts the number of objects in images (Figure 7).

We next investigate whether language model guidance can elicit better captions from a model trained only on

a photo of two dogs   a photo of one bird   a photo of four suitcases   a photo of five sheep

Figure 7. Captions generated with LM guidance with a prompt of 25 captions in the form of "a photo of NUMBER OBJECTS". Examples are selected to show different numbers of objects.
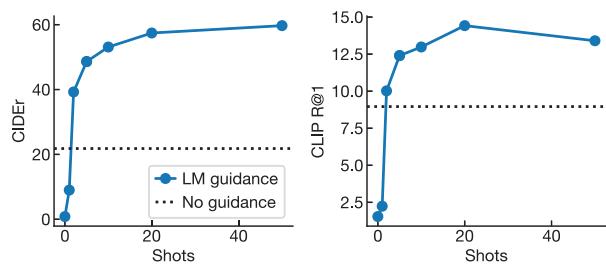

Figure 8. LM guidance substantially improves CIDEr and retrieval scores of a model trained solely on minimally curated web data and evaluated on MS-COCO. The x-axis shows the number of captions used to prompt the LM; we do not prompt with images.

low-quality data. Here, we use a CoCa model that is pretrained on image-alt text pairs from the web (the ALIGN dataset [17]) and classification labels converted to text (the JFT-5B dataset [49]), without any additional fine-tuning. Because the data distribution places higher probability mass on short, non-descriptive captions than on longer captions, the resulting model is of limited utility for captioning, and would generally need to be fine-tuned on another dataset such as MS-COCO before being applied to a captioning task. Rather than fine-tune, we use LM guidance to prompt the model with captions from the MS-COCO training set.

LM guidance substantially improves the quality of the captions produced by the original pretrained CoCa model without any clean parallel data. With LM guidance, we achieve CIDEr scores of 48.6 with 5 shots and 59.7 with 50 shots, far exceeding the CIDEr score of 21.8 obtained with no guidance. Figure 8 shows CIDEr and CLIP recall@1 scores for LM guidance of this pretrained CoCa model as a function of the number of shots, with $\alpha = \beta = 1$. Table 3 compares classifier-free guidance and LM guidance. CFG yields higher CLIP-Scores and retrieval accuracy than LM guidance with $\alpha = \beta = 1$, but LM guidance provides much higher CIDEr scores.

We compare captions generated with CFG to those generated with LM guidance for four images in Figure 9. In general, CFG produces agrammatical captions, whereas LM guidance produces grammatical captions but hallucinates details. For example, the image in the upper left shows two elephants and no zebras, but LM guidance leads to the caption "an elephant and a zebra in a field."

	Reference-Based Metrics					Reference-Free Metrics				
Model	BLEU-4	METEOR	ROUGE	CIDEr	RefOnlyCLIP-S	CLIP-S	R@1	R@5	R@10	RefCLIP-S
Classifier-free guidance:										
$\gamma = 1.0$	8.2	8.3	21.8	21.8	0.766	0.694	9.0	19.5	26.1	0.725
$\gamma = 1.2$	8.6	9.5	24.5	25.0	0.781	0.718	12.7	27.2	35.1	0.745
$\gamma = 1.5$	8.9	10.0	25.6	25.2	0.780	0.728	16.7	33.8	43.0	0.750
$\gamma = 2.0$	8.1	9.7	23.9	22.9	0.777	0.741	21.2	40.8	51.1	0.755
$\gamma = 3.0$	7.1	8.7	20.0	18.5	0.767	**0.753**	25.8	47.8	58.3	0.756
$\gamma = 4.0$	6.4	7.5	16.3	13.9	0.749	0.743	**27.3**	**48.5**	**58.1**	0.742
Language model guidance with $\alpha = \beta = 1$:										
2 captions	12.7	14.6	34.7	39.3	0.806	0.688	10.0	23.7	32.4	0.740
5 captions	15.0	16.6	39.1	48.6	0.827	0.712	12.4	27.5	37.5	0.763
10 captions	16.2	17.7	40.5	53.1	0.835	0.723	13.0	30.5	41.0	0.773
20 captions	17.4	18.4	41.6	57.4	0.839	0.728	14.4	32.2	42.7	0.777
50 captions	**18.1**	**19.1**	**42.5**	**59.7**	**0.840**	0.729	13.4	32.5	43.9	**0.778**
Other models trained without aligned MS-COCO images and captions:										
ZeroCap [41]	2.6	11.5		14.6		0.87				0.79
MAGIC [39]	12.9	17.4	39.9	49.3						
Flamingo [3]				84.3						
DeCap (560 captions) [22]				51.4						
DeCap (full train set) [22]	24.7	25.0		91.2						
CapDec (full train set) [27]	26.4	25.1	51.8	91.8						

Table 3. Comparison of decoding strategies for a captioning model trained only on minimally curated web data (JFT-5B and ALIGN) and evaluated on MS-COCO. At the bottom, we report metrics for other models trained without aligned MS-COCO images and captions. These models may not be directly comparable since they use different pretraining data. DeCap and CapDec use all 560K captions in the MS-COCO training set to train their decoders; we include CIDEr for DeCap with 560 captions (0.1% of the training data) for comparison.

γ=1.0: a photo of the small elephant.
γ=2.0: elephants in the ruaha national park
γ=3.0: elephants chobe np
LM: a elephant and a zebra in a field
GT: Two elephants standing on a grassy field next to a tree.

γ=1.0: a photo of the small coffee.
γ=2.0: a coffee in a video game.
γ=3.0: a banana in a video game.
LM: a banana with a cup of coffee
GT: A close up of a banana next to a cup with liquid.

γ=1.0: a photo of the large windsurfing.
γ=2.0: windsurfing in tarifa
γ=3.0: windsurfing wallpapers 1200x1024
LM: a windsurfer in the water
GT: A man riding a wind sail in the ocean filled with waves.

γ=1.0: a photo of the large giraffe.
γ=2.0: a giraffe in a video game.
γ=3.0: giraffe standing photo 1
LM: a giraffe standing in a tall tree
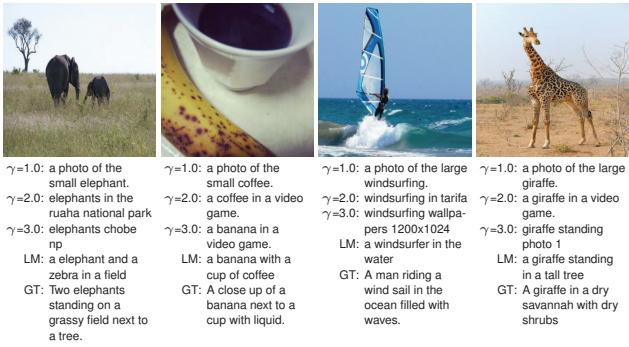GT: A giraffe in a dry savannah with dry shrubs

Figure 9. Examples of captions generated from a model pretrained only on minimally curated data, for randomly selected images without human faces. Captions labeled $\gamma = 1.0$ are obtained without CFG; $\gamma > 1$ uses CFG; LM indicates LM guidance with $\alpha = \beta = 1$ and 20 shots; GT indicates ground truth.

5. Conclusion

Our study indicates that it is possible to substantially improve the extent to which generated captions uniquely describe the images they correspond to, raising questions regarding the goal of image captioning and how it should be evaluated. As it is conventionally formulated, image captioning aims not to provide text that can substitute for an image, but to write the text that a human annotator would have written. This formulation penalizes captions that are more descriptive than ground truth, even when a human might prefer them. On the other hand, treating image captioning as a problem of generating a caption that lies close to the image in the embedding space of an image-text model is also inadequate, because captions that lie close to the image need not be grammatical and may contain gibberish. Our proposed methods leveraging classifier-free guidance and language model guidance modulate the trade-offs between these two goals, as captured by various reference-based and reference-free metrics.

There are several possible extensions to our work. First, our present experiments use only greedy decoding. Although greedy decoding appears to perform reasonably well in our setup, it may be suboptimal for LM guidance with prompts that impose greater structure upon the captions. If the LM is prompted to output either "there is a person in this image" or "there is no person this image", greedy decoding is likely to fail even if the captioner properly scores the two possible captions, because when choosing between the tokens "a" and "no", the captioner has no knowledge of the structure that the LM will impose on future tokens. Since beam search could explore both tokens, it may offer better results in this scenario. Second, our method could be combined with RL-based methods to increase similarity in a contrastive embedding space, which may further improve retrieval performance and CLIPScore. Finally, with a perfect captioning model, $p(\text{image}|\text{caption})$ should increase with γ. However, in practice we find that $\gamma > 3$ leads to a decrease in retrieval performance. This discrepancy suggests that the difference between the conditional and unconditional model distributions may be a noisy estimator of the pointwise mutual information. Although selecting γ is one way to regularize this estimator, there may also be strategies to regularize $p(\text{x}|\text{y})/p(\text{x})$ at training time.

Acknowledgements

References

[1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022. 2

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019. 6

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 2, 8

[4] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4662–4670, 2022. 5

[5] David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. ic^3: Image captioning by committee consensus. *arXiv preprint arXiv:2302.01328*, 2023. 2

[6] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, Seattle, United States, July 2022. Association for Computational Linguistics. 2, 5

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 4

[8] Katherine Crowson. You can apply a similar trick to classifier-free guidance to autoregressive transformers to sample from a synthetic "super-conditioned" distribution. https://twitter.com/RiversHaveWings/status/1478093658716966912, 2022. 3

[9] Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Ph. D. thesis, Université Paris 7, 1994. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. 3

[11] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015. 2

[12] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 2, 3

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2, 4

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 3

[15] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 6

[16] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1138–1147, 2021. 2, 5

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 4, 7

[18] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015. 4

[19] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv preprint arXiv:2112.04139*, 2021. 5

[20] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. Transparent human evaluation for image captioning. *arXiv preprint arXiv:2111.08940*, 2021. 1, 2

[21] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. 6

[22] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning. In *International Conference on Learning Representations*, 2022. 2, 8

[23] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[25] Annika Lindh, Robert J Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D Kelleher. Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks*, pages 176–187. Springer, 2018. 2

[26] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[27] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 2, 8

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 1

[29] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[34] Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models to fuse visual data into enriched image captions. *arXiv preprint arXiv:2305.17718*, 2023. 2

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*. 2

[36] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023. 3

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018. 4, 6

[38] David R So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Primer: Searching for efficient transformers for language modeling. *arXiv preprint arXiv:2109.08668*, 2021. 4

[39] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 2, 8

[40] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 4

[41] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 2, 5, 8

[42] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[43] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[44] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023. 2

[45] Shitong Xu. Clip-diffusion-lm: Apply diffusion model on image captioning. *arXiv preprint arXiv:2210.04559*, 2022. 2

[46] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017. 2

[47] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3, 4, 6, 14

[48] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yin-

fei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2, 3

[49] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 4, 7

[50] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 5

[51] Youyuan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. Distincive image captioning via clip guided group optimization. *arXiv preprint arXiv:2208.04254*, 2022. 2, 5

[52] Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. Exploring discrete diffusion models for image captioning. *arXiv preprint arXiv:2211.11694*, 2022. 2