Evaluating Cross-modal Generative Models Using Retrieval Task

Shivangi Bithel
shivangi.bithel@cse.iitd.ac.in
IIT Delhi
New Delhi, India

Srikanta Bedathur
srikanta@cse.iitd.ac.in
IIT Delhi
New Delhi, India

ABSTRACT

Generative models have taken the world by storm – image generative models such as Stable Diffusion and DALL-E generate photo-realistic images, whereas image captioning models such as BLIP, GIT, ClipCap, and ViT-GPT2 generate descriptive and informative captions. While it may be true that these models produce remarkable results, their systematic evaluation is missing, making it hard to advance the research further. Currently, heuristic metrics such as the Inception Score and the Fréchet Inception Distance are the most prevalent metrics for the image generation task, while BLEU, CIDEr, SPICE, METEOR, BERTScore, and CLIPScore are common for the image captioning task. Unfortunately, these are poorly interpretable and are not based on the solid user-behavior model that the Information Retrieval community has worked towards. In this paper, we present a novel cross-modal retrieval framework to evaluate the effectiveness of cross-modal (image-to-text and text-to-image) generative models using reference text and images. We propose the use of scoring models based on user behavior, such as Normalized Discounted Cumulative Gain (*nDCG'@K*) and Rank-Biased Precision (*RBP'@K*) adjusted for incomplete judgments. Experiments using ECCV Caption and Flickr8k-EXPERTS benchmark datasets demonstrate the effectiveness of various image captioning and image generation models for the proposed retrieval task. Results also indicate that the nDCG'@K and RBP'@K scores are consistent with heuristics-driven metrics, excluding CLIPScore, in model selection.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Evaluation of retrieval results**; **Relevance assessment**.

KEYWORDS

Evaluation method; cross-modal generative modal; cross-modal retrieval

ACM Reference Format:
Shivangi Bithel and Srikanta Bedathur. 2023. Evaluating Cross-modal Generative Models Using Retrieval Task. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Cross-modal generative models such as stable-diffusion –and its many variants– and DALL-E 2 (for text-to-image generation); ClipCap, BLIP, ViT-GPT2, GIT, etc. (for image caption generation), have become very popular in recent times. All these models are trained to learn the alignment between image and text [24, 39] over billion-scale datasets such as LAION5B [44] collected by crawling image-caption pairs from the internet. While these models have demonstrated remarkable user experience, their systematic evaluation is still in its nascent stages. Cross-modal (image-to-text and text-to-image) generation task is used to answer the question whether these models are learning meaningful associations between the two modalities [23]. The evaluation is expected to reveal whether text-to-image generative models generate a similar image based on the semantics of textual description, whereas image-to-text generative models describe the semantic content of the image using meaningful and descriptive natural language.

Unfortunately, the current systematic evaluation of these models suffers from a number of critical issues:

Poor Interpretability of Metrics Commonly used metrics for evaluating image-to-text generative models are based on the co-occurrence frequency of n-grams in the predicted caption and the human written reference caption [7, 25, 32, 47], or on the text distance between the generated caption and reference caption [50], or on the embedding distance between the generated caption and the input image [15]. Similarly, the text-conditioned image generation models are either evaluated using the divergence between the conditional class distribution and the marginal class distribution of the generated image and generated dataset respectively [8] or using the difference of two Gaussians fitted to the real-world and generated image data measured using Fréchet distance [16]. The quality of these metrics highly depends on the features returned by the inception net [9]. Moreover, these metrics are not robust to new words and favor familiar words and the style of the captions. These issues are illustrated in Figure 1 using an example from the Flickr8k-EXPERTS dataset. The BLEU-4 score for all the generated captions in the example is around 0 as there is no 4-gram overlap with the human-written reference captions. Moreover, the BLIP method and Stable Diffusion V2 generate the most semantically aligned caption and image, however, the caption with repeated words and the image with less photorealism receives the highest CLIPScore. Thus, it becomes difficult to interpret the scores generated by these metrics.

Lack of a User-behavior Model Information retrieval community has stressed the importance of having a realistic user-behavior model while developing evaluation metrics for ranked results [28]. For instance, Discounted Cumulative Gain (DCG) considers the model of a user who inspects the results in ranked order, with exponentially discounted satisfaction as she goes down the rankings [18].

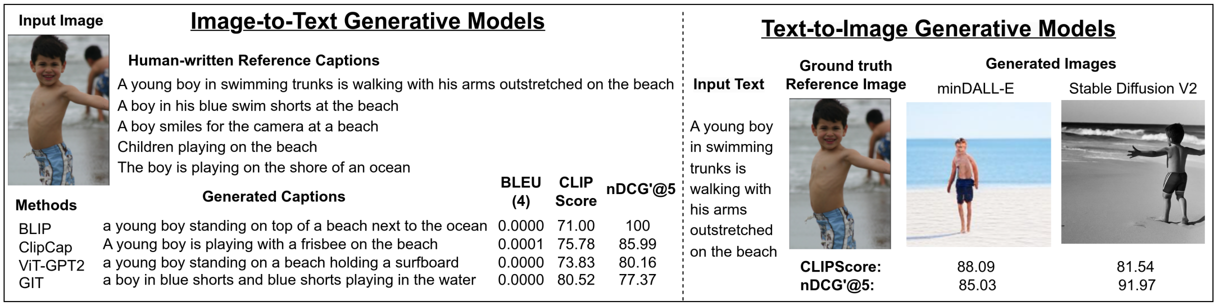

Figure 1: Example showing shortcomings of well-known n-gram matching and embedding-based metrics

More recent models such as Rank-Biased Precision (RBP) [28] consider slightly more sophisticated (and realistic) user models. However, none of the metrics used in the evaluation of cross-modal retrieval have a well-defined user-behavior model. Similar criticism on the choice of metrics used by these models has also been made by Musgrave et al. [30] and Chun et al. [11].

Extremely Shallow Judgments Musgrave et al., [30] also recommends the use of Mean Average Precision at R (MAP@R) to compare the ranking of retrieved results. Unfortunately, even this recommendation (also used by Chun et al. [11]), is fraught with problems when the human judgments are conducted to a very shallow depth leading to the retrieved ranking containing incomplete judgments [28, 42]. We found that two of the most commonly used human-judgment datasets, *viz.,* ECCV Caption [11] and Flickr8k-EXPERTS [17], contain incomplete relevance assessments and very shallow depth of judgments (average evaluation depth of 12 and 6 on ECCV and Flickr8k-EXPERTS respectively), making the choice of MAP@R questionable.

1.1 Contributions

In this paper, we investigate whether the heuristics-based metrics such as CLIPScore, BLEU-4, FID, etc. used for evaluating cross-modal generative models are consistent with systematic metrics for ranked retrieval evaluation such as nDCG'@K [43] and RBP'@K [28]. For this purpose, we propose a novel unified cross-modal retrieval (CMR) framework that computes a ranking of results for a given query by making use of a cross-modal generative model (Section 3). We conduct an experimental evaluation using ECCV Caption and Flickr8k-EXPERTS benchmarks which contain graded (albeit shallow) relevance assessments (Section 4). Our results indicate that although CLIPScore score trends seem to be consistent with nDCG' and RBP' scores for Text-to-Image models, this is not the case for Image-to-Text (captioning) models. Further, we observe that there is a bigger spread of CLIPScore values for different captioning models on the Flickr8k-EXPERTS dataset and FID Scores for the two image generation models on both the datasets, than nDCG'@K and RBP'@K scores (Section 5).

2 BACKGROUND

Multimodal learning has grown rapidly in recent years with pre-trained vision-language models [20, 34]. Image-to-text generation, also known as image captioning has made significant progress in generating captions that are indistinguishable from those written by humans. The task uses an image as input and generates its natural language description. Some of the captioning models, including BLIP [24], and GIT [48] are generative unified transformer frameworks that have been trained on multiple tasks involving different modalities, whereas others such as ClipCap [29], MAPL [27], and FROZEN [46] have only been trained on the image captioning task.

Various reference-based, reference-free, and self-retrieval-based methods are used to evaluate and compare the effectiveness of image captioning models in generating valid and descriptive captions for a given image. The majority of these reference-based evaluation metrics, such as BLEU [32], CIDEr [47], METEOR [7], and ROUGE [25], investigate the co-occurrence frequency of n-grams in the predicted caption in comparison to five human-written reference captions, whereas methods like SPICE [6] apply a semantic parser to a set of references and compute similarity using the predicted scene graph. Popular embedding-based metrics, such as BERTScore [50], employ contextual embeddings to represent tokens and compute matching using cosine similarity, optionally weighted with inverse document frequency scores. CLIPScore [15], a popular reference-free evaluation metric, computes the cosine similarity between features extracted from the image and candidate caption using CLIP's feature extractor. The self-retrieval-based evaluation ranks the set of original images using the generated caption as the query to produce a ranked list. It computes the top-k recalls based on the ranked lists, which is the proportion of images within the top-k positions of the ranked lists for each query. The top-k recall is an excellent indicator of how well a model captures distinctiveness in its descriptions. Our proposed text-to-image retrieval task is a combination of reference-based and self-retrieval-based methods, favoring the generation of semantically relevant and unique captions in its evaluations.

Text-to-image generation, also known as image generation has also made significant progress in generating high-quality photo-realistic images from a given text prompt. These models are mainly divided into four groups, namely normalizing flows [38], VAE [21], GAN [14] and diffusion models [36, 39, 40]. The task uses an input text prompt to generate a semantically similar image from the latent space. Some of the recent models are diffusion-based models, which include DALL-E[37], DALL-E 2[36], minDALL-E [19], Stable Diffusion [39], GLIDE [31], Make-A-Scene [13], and IMAGEN [40]. To evaluate and compare these implicit image-generative models, we require an empirical measure. The most common metrics used are Inception Score (IS), Fréchet Inception Distance (FID), and Fréchet Clip Distance (FCD) [9]. The IS uses an Inception-v3 Network pre-trained on ImageNet and calculates a statistic of the network's outputs when applied to generated images. FID computes Fréchet Distance between two multivariate Gaussians, fitted to the features extracted by the inception network at pool3 layer for real and generated data. As these metrics are based on features

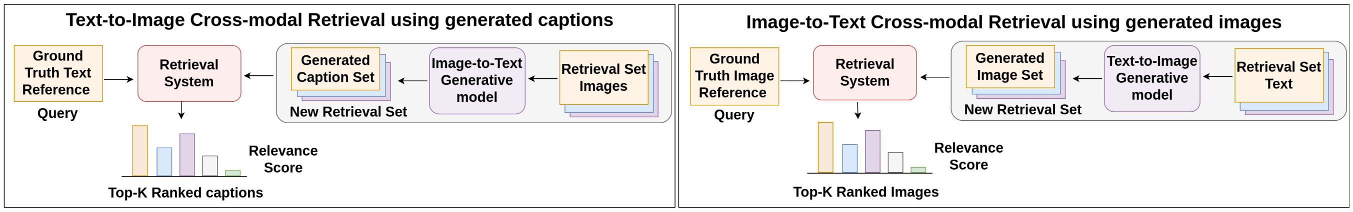

Figure 2: Cross-modal Retrieval Framework to Evaluate Generative Models.

and scores computed using a pre-trained network on ImageNet [22] dataset with a particular image size, it is not clear how well they transfer to other image types and image sizes. In contrast to the heuristic-based metric, our proposed retrieval task models the user behavior and judgment of relevant items. It also uses the pre-trained transformer-based models trained on large-scale data to provide robustness to different data types and sizes.

3 CMR FRAMEWORK FOR EVALUATION

In this section, we describe the image captioning task and its representative models, the image generation task and its representative models, as well as our proposed cross-modal retrieval framework for evaluating the cross-modal generative models. The CMR framework, as depicted in Figure 2, consists of a set of human-written captions and real-world images as ground truth reference queries, a retrieval system, a set of generated items (captions and images) as new retrieval set, a generative model for evaluation, an input retrieval set of images and textual prompt, and a set of evaluation metrics (nDCG'@K and RBP'@K).

3.1 Caption Generation Models

For a given image $x \in \mathbb{R}^{H \times W \times C}$ of height H, width W, color channels C, the captioning task concerns generating a description y consisting of words from the vocabulary of natural language. We use four representative caption generation models BLIP, GIT, ClipCap, and ViT-GPT2 in our experiments. The BLIP model consists of a transformer encoder [12] to comprehend visual features and an image-grounded text decoder to generate auto-regressive text. The GIT model uses a contrastive pre-trained model as an image encoder [49] and a transformer text decoder. The ClipCap employs a pre-trained CLIP [34] network as an image encoder and a pre-trained language model, GPT-2 [35], as a text generator. A transformer-based mapping network predicts caption tokens conditioned on the prefix in an autoregressive manner. We also use the ViT-GPT2 model which uses ViT as an image encoder and GPT2 as a text decoder. All the models are pre-trained on image-text pairs and finetuned for the captioning task on the MS-COCO [26] dataset.

3.2 Image Generation Models

For a given natural language description y, the image generation task concerns generating an image $x \in \mathbb{R}^{H \times W \times C}$ of height H, width W, color channels C. We use two representative image generation models, minDALL-E, and Stable Diffusion V2 in our experiments. MinDALL-E is a small, easily accessible two-stage autoregressive model based on DALL-E. In stage-1, minDALL-E generates high-quality image samples using VQGAN, and in stage-2, it uses a decoder-only sparse transformer trained from scratch on 14 million image-text pairs from CC3M [45] and CC12M [10]. In addition, we employ the latent diffusion model known as Stable Diffusion V2 (SDV2). It consists of a text encoder, a variational autoencoder (VAE), and a U-Net. The SDV2 model has been pre-trained on the LAION-5B dataset [44].

3.3 Evaluation of Image Captioning Models

The evaluation of the image captioning model can be framed as a text-to-image retrieval task as shown on the left side of Figure 2. This procedure encourages the image captioning model to generate a semantically similar caption for the input image that not only describes the image scene but is also unique in its description. A semantically similar caption will be able to rank the human-annotated relevant images at the top when the ground truth reference caption is used as the query. First, we feed the set of input images X, one by one to the caption generation model, say BLIP, to generate a single caption candidate y. The generated candidate caption set Y forms the new retrieval set used as an intermediary for the cross-modal retrieval task. Then we perform the text-to-image retrieval task, using the corresponding ground truth reference caption y_x as the query to rank the generated caption set Y and its mapped image set X using cosine similarity score. Finally, we evaluate the ranking of the top-k images based on the available human-annotated graded relevance score using nDCG'@K and RBP'@K metrics. The higher the ranking score, the better the image captioning model.

3.4 Evaluation of Image Generation Models

Figure 2 on the right shows the reverse procedure of the evaluation of the image generative model, framed as an image-to-text retrieval task. This procedure encourages the image generation model to generate a semantically similar image that is not only representative of the input textual prompt but also photo-realistic and unique. A semantically similar image will be able to rank the human-annotated relevant textual prompts at the top when the ground truth real image is used as a query. First, we feed the set of input textual prompts Y, one by one to the text-conditioned image synthesis model, say minDALL-E, to generate a single image candidate x. The generated candidate image set X forms the new retrieval set used as an intermediary for the cross-modal retrieval task. Then we perform the image-to-text retrieval task, using the corresponding ground truth reference image x_y as the query to rank the generated image set X and its mapped textual prompt set Y using cosine similarity score. Finally, we evaluate the ranking of the top-k textual prompt based on the available human-annotated graded relevance score using nDCG'@K and RBP'@K metrics. The higher the ranking score, the better the image generation model.

4 EXPERIMENTAL SETUP

Datasets: *ECCV Caption* [11] and *Flickr8k-EXPERTS* datasets are extended subsets of the COCO Caption [26] and Flickr-8K datasets [17] respectively. ECCV Caption includes 1,332 query images and 1,261 query captions, while Flickr8k-EXPERTS includes 1,000 images and 977 captions. The dataset contains a rating score of image-caption pairs given by human experts on a scale of 0 to 3, with

Table 1: Evaluation of Image-to-Text Generative Model on ECCV Caption and Flickr8k-EXPERTS Dataset. Bold fonts and underline indicate the best performer and the second-best performer respectively. Results marked † are statistically significant (i.e., two-sided t-test with $p \leq 0.05$) over the second-best method.

Dataset	Method	nDCG'@5	nDCG'@10	nDCG'@15	RBP'@5	RBP'@10	RBP'@15	BLEU-4 ↑	CLIPScore ↑	CLIPRefScore ↑
ECCV Caption	ClipCap	80.46	87.96	91.25	1.47	1.87	1.91	33.3	77.6	82.3
ECCV Caption	ViT-GPT2	81.11	88.4	91.54	1.48	1.88	1.92	39.2	75.4	81.8
ECCV Caption	GIT	81.75	88.74	91.91	1.49	1.88	1.93	37.8	**77.9†**	83.2
ECCV Caption	BLIP	**82.35†**	**89.15†**	**92.24†**	**1.5†**	**1.89†**	**1.94†**	**41.7†**	77.8	**83.4†**
Flickr8k-EXPERTS	ClipCap	68.38	70.23	70.39	0.84	0.85	0.85	18.69	77.55	79.11
Flickr8k-EXPERTS	ViT-GPT2	67.27	69.32	69.45	0.83	0.83	0.83	25.68	**78.53†**	81.41
Flickr8k-EXPERTS	GIT	69.46	71.08	71.23	0.86	0.87	0.87	17.21	71.47	75.22
Flickr8k-EXPERTS	BLIP	**69.83**	**71.44**	**71.51**	**0.87†**	**0.88†**	**0.88†**	**29.14†**	77.81	**81.52†**

Table 2: Evaluation of Text-to-Image Generative Model on ECCV Caption and Flickr8k-EXPERTS Dataset. Bold fonts indicate the best performer method. Results marked † are statistically significant (i.e., two-sided t-test with $p \leq 0.05$) over the second-best performer.

Dataset	Method	nDCG'@5	nDCG'@10	nDCG'@15	RBP'@5	RBP'@10	RBP'@15	FID ↓	CLIPScore ↑	FCD ↓
ECCV Caption	MinDALL-E	73.93	77.85	83.5	2.10	2.16	2.16	50.61	78.68	20.31
ECCV Caption	SDV2	**77.9†**	**81.02†**	**86.04†**	**2.22†**	**2.28†**	**2.29†**	**18.31**	**83.07†**	**13.59**
Flickr8k-EXPERTS	MinDALL-E	73	75.02	75.02	0.90	0.90	0.90	99.99	79.66	24.69
Flickr8k-EXPERTS	SDV2	**75.74†**	**77.03†**	**77.03†**	**0.95†**	**0.95†**	**0.95†**	**63.38**	**85.88†**	**14.97**

0 indicating that the caption does not describe the image at all, 1 indicating that the caption describes minor aspects of the image but does not describe the image, 2 indicating that the caption almost describes the image with minor errors, and 3 indicating that the caption describes the image.

Implementation Details: For image captioning, we use the open implementation of ClipCap [2], ViT-GPT2 [5], GIT [3] and BLIP [1] model. For image generation, we use the open implementation of Stable Diffusion V2 [4] and minDALL-E [19] as DALL-E and DALL-E 2 are not freely accessible for research purposes. All models are taken from the HuggingFace library and Github, without any further fine-tuning. To extract the image and text features, we used Swin-Large Transformer Encoder and SBERT (`distilroberta-base`) respectively. For a fair comparison, we used the best sampling settings provided for each model and a seed of 3407 [33] to generate the captions and the images. We used cosine distance to measure similarity.

Evaluation Metrics: We propose to use user-model-based judgment metrics namely Normalized Discounted Cumulative Gain (nDCG') [41] and Rank Biased Precision (RBP') [28] adjusted for incomplete judgments to evaluate our CMR framework. In our experiments, we used a condensed list in Qrels, and removed all the unjudged documents from the ranking, to compute *nDCG'* and *RBP'*. The *nDCG* value for top-K retrieved elements is expressed as $nDCG@K = \frac{DCG@K}{IDCG@K}$ where *DCG* and *IDCG* are the Discounted Cumulative Gain and Ideal Discounted Cumulative Gain. RBP is based on the monotonically decreasing values in a geometric sequence. It can be expressed as, $RBP(R,p) = (1-p)\sum_{i=1}^{|R|} r_i p^{i-1}$ where p is an abstraction of the user's searching persistence, expressed between 0 and 1, R represents the relevance vector to be evaluated, and r_i indicates the relevance of the document ranked in position i within the ranking. We use $p = 0.5$ to account for shallow judgments as recommended by the authors [28].

5 EXPERIMENTAL RESULTS

We address these two research questions in our experiments:

RQ1: What is the ranking effectiveness of generative models in a Cross-modal Retrieval (CMR) task?

RQ2: Are heuristics-driven metrics used in generative model evaluation consistent with results from user-behavior-driven metrics such as *nDCG'@K* and *RBP'@K*?

In Table 1, nDCG'@K and RBP'@K compare the performance of image-captioning models on ECCV Caption and Flickr8k-EXPERTS. The BLIP model gets the highest nDCG'@K and RBP'@K scores in comparison to other models for both datasets. The same model also outperforms others in heuristics-driven metrics as well. This suggests that BLIP captions are not only semantically more similar to the ground truth reference captions but also can better rank images in the retrieval task. This is also evident from the example in Figure 1. With respect to RQ2, we notice that CLIPScore is inconsistent with ranking metrics for different models – the GIT and ViT-GPT2 models get the highest CLIPScore on the ECCV Caption and Flickr8-EXPERTS respectively. It is also interesting to note that there is a bigger spread of CLIPScore values for different models on the Flickr8k-EXPERTS dataset, than nDCG'@K and RBP'@K scores. The images generated by the Stable Diffusion V2 (SDV2) and minDALL-E models are compared for their ranking effectiveness in Table 2. The results suggest that the SDV2 model generates a more similar image for a given text input that is also distinct from the set of generated images in order to rank textual prompts more effectively in the retrieval task for both datasets. Also, there is a huge gap in FID Score for the two models, while it is not the case with nDCG'@K and RBP'@K scores.

6 CONCLUSION

In this paper, we explored whether heuristics-based metrics used for evaluating image-to-text and text-to-image generative models are consistent with models such as nDCG'@K and RBP'@K that are based on robust user behavior models. We presented a novel unified cross-modal retrieval framework that uses generative models for the retrieval task and used it in our comparison of metrics. Empirically we showed the interpretability challenge with the heuristics metrics and showed that nDCG'@K and RBP'@K are more suitable in terms of their interpretability and usability. Further investigation is needed to use the nDCG'@K and RBP'@K metrics to tune the underlying models, and also to develop better evaluation benchmarks with graded judgments further deep in rankings.

REFERENCES

[1] 2022. BLIP Model. https://github.com/salesforce/BLIP.

[2] 2022. CLIPCAP Model. https://github.com/rmokady/CLIP_prefix_caption.

[3] 2022. GIT Model. https://huggingface.co/microsoft/git-base-coco.

[4] 2022. Stable Diffusion Model. https://huggingface.co/stabilityai/stable-diffusion-2-1.

[5] 2022. ViT-GPT2 Model. https://huggingface.co/nlpconnect/vit-gpt2-image-captioning.

[6] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 9909)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 382–398. https://doi.org/10.1007/978-3-319-46454-1_24

[7] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. https://aclanthology.org/W05-0909/

[8] Shane Barratt and Rishi Sharma. 2018. A Note on the Inception Score. arXiv:1801.01973 [stat.ML]

[9] Eyal Betzalel, Coby Penso, Aviv Navon, and Ethan Fetaya. 2022. A Study on the Evaluation of Generative Models. *CoRR* abs/2206.10935 (2022). https://doi.org/10.48550/arXiv.2206.10935 arXiv:2206.10935

[10] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. (2021), 3558–3568. https://doi.org/10.1109/CVPR46437.2021.00356

[11] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. 2022. ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII (Lecture Notes in Computer Science, Vol. 13668)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 1–19. https://doi.org/10.1007/978-3-031-20074-8_1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV (Lecture Notes in Computer Science, Vol. 13675)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 89–106. https://doi.org/10.1007/978-3-031-19784-0_6

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 7514–7528. https://doi.org/10.18653/v1/2021.emnlp-main.595

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6626–6637. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

[17] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Intell. Res.* 47 (2013), 853–899. https://doi.org/10.1613/jair.3994

[18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[19] Saehoon Kim, Sanghun Cho, Chiheon Kim, Doyup Lee, and Woonhyuk Baek. 2021. minDALL-E on Conceptual Captions. https://github.com/kakaobrain/minDALL-E.

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5583–5594. http://proceedings.mlr.press/v139/kim21k.html

[21] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[23] Hang Li, Jindong Gu, Rajat Koner, Sahand Sharifzadeh, and Volker Tresp. 2022. Do DALL-E and Flamingo Understand Each Other? *CoRR* abs/2212.12249 (2022). https://doi.org/10.48550/arXiv.2212.12249 arXiv:2212.12249

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

[25] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[27] Oscar Mañas, Pau Rodríguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2022. MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting. *CoRR* abs/2210.07179 (2022). https://doi.org/10.48550/arXiv.2210.07179 arXiv:2210.07179

[28] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (2008), 2:1–2:27. https://doi.org/10.1145/1416950.1416952

[29] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *CoRR* abs/2111.09734 (2021). arXiv:2111.09734 https://arxiv.org/abs/2111.09734

[30] Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. A Metric Learning Reality Check. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV (Lecture Notes in Computer Science, Vol. 12370)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 681–699. https://doi.org/10.1007/978-3-030-58595-2_41

[31] 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, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 16784–16804. https://proceedings.mlr.press/v162/nichol22a.html

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[33] David Picard. 2021. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *CoRR* abs/2109.08203 (2021). arXiv:2109.08203 https://arxiv.org/abs/2109.08203

[34] 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. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang

(Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html
[35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* abs/2204.06125 (2022). https://doi.org/10.48550/arXiv.2204.06125 arXiv:2204.06125
[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. http://proceedings.mlr.press/v139/ramesh21a.html
[38] Danilo Jimenez Rezende and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 1530–1538. http://proceedings.mlr.press/v37/rezende15.html
[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042
[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR* abs/2205.11487 (2022). https://doi.org/10.48550/arXiv.2205.11487 arXiv:2205.11487
[41] Tetsuya Sakai. 2007. Alternatives to Bpref. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando (Eds.). ACM, 71–78. https://doi.org/10.1145/1277741.1277756
[42] Tetsuya Sakai. 2021. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1, Article 12 (feb 2021), 8 pages. https://doi.org/10.1145/3451964.3451976
[43] Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr.* 11, 5 (2008), 447–470. https://doi.org/10.1007/s10791-008-9059-7
[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *CoRR* abs/2210.08402 (2022). https://doi.org/10.48550/arXiv.2210.08402 arXiv:2210.08402
[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2556–2565. https://doi.org/10.18653/v1/P18-1238
[46] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 200–212. https://proceedings.neurips.cc/paper/2021/hash/01b7575c38dac42f3cfb7d500438b875-Abstract.html
[47] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087
[48] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. https://doi.org/10.48550/ARXIV.2205.14100
[49] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A New Foundation Model for Computer Vision. *CoRR* abs/2111.11432 (2021). arXiv:2111.11432 https://arxiv.org/abs/2111.11432
[50] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr