Do DALL-E and Flamingo Understand Each Other?

Hang Li*[1,2] Jindong Gu*[3] Rajat Koner[1] Sahand Sharifzadeh[1] Volker Tresp[1,2]
[1]LMU Munich, Germany [2]Siemens AG, Germany [3]University of Oxford, UK

Abstract

The field of multimodal research focusing on the comprehension and creation of both images and text has witnessed significant strides. This progress is exemplified by the emergence of sophisticated models dedicated to image captioning at scale, such as the notable Flamingo model and text-to-image generative models, with DALL-E serving as a prominent example. An interesting question worth exploring in this domain is whether Flamingo and DALL-E understand each other. To study this question, we propose a reconstruction task where Flamingo generates a description for a given image and DALL-E uses this description as input to synthesize a new image. We argue that these models understand each other if the generated image is similar to the given image. Specifically, we study the relationship between the quality of the image reconstruction and that of the text generation. We find that an optimal description of an image is one that gives rise to a generated image similar to the original one. The finding motivates us to propose a unified framework to finetune the text-to-image and image-to-text models. Concretely, the reconstruction part forms a regularization loss to guide the tuning of the models. Extensive experiments on multiple datasets with different image captioning and image generation models validate our findings and demonstrate the effectiveness of our proposed unified framework. As DALL-E and Flamingo are not publicly available, we use Stable Diffusion and BLIP in the remaining work. Project website: https://dalleflamingo.github.io.

1. Introduction

Recently, multimodal research that aims to improve machine understanding of images and text has made significant advances [42, 43, 46, 8, 38, 59, 67, 21]. Text-to-image generation models such as DALL-E [44, 43] and Stable Diffusion (SD) [46] are capable of converting complex textual descriptions [47] from real-world scenarios into high-fidelity images [43, 24, 40, 44]. Conversely, image-to-text generation models, e.g., Flamingo [2] and BLIP [35], exhibit the ability to comprehend the intricate semantics present in images and produce coherent descriptions [57, 26, 56, 55, 62, 37, 60]. Despite the closeness of the image captioning and text-to-image generation tasks, they are often studied in isolation from each other, i.e., the communication between these models is underexplored [38, 30]. This raises an interesting question: do image-to-text generation models and text-to-image generation models possess mutual understanding? Concretely, we investigate this question by letting an image-to-text model, BLIP, generate a text description for a given image, which subsequently serves as input to a text-to-image model, SD, to synthesize a new image[1]. We argue that BLIP and SD understand each other if the generated image is similar to the source image. Such mutual understanding may enhance their respective abilities to comprehend underlying concepts, resulting in superior caption generation and image synthesis. Figure 1 illustrates this idea, where the upper caption is a better representation of the input image than the

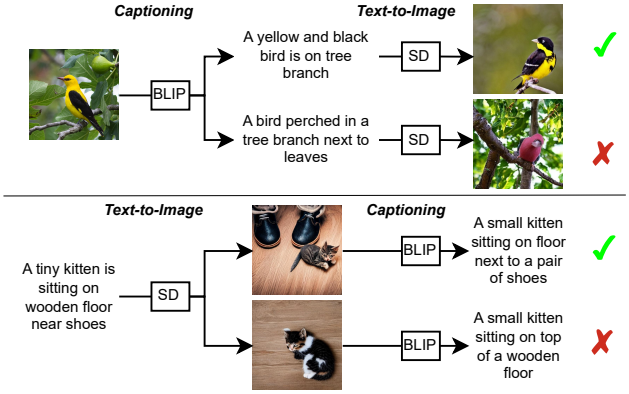

Figure 1. Illustration of the communication tasks between SD and BLIP. Top: SD generates an image for each caption created by BLIP, where an accurately reconstructed image indicates a more precise caption. Bottom: The reverse task involves SD generating image candidates, which are then used by BLIP to produce captions for those images. The best image that represents a text is the one that leads to the best reconstruction of the original text.

*Equal contribution. Correspondence to ha.li@campus.lmu.de

[1]The initial idea of this work was motivated by Flamingo and DALL-E. However, their model weights are unavailable at the time of publication.

lower caption as it leads to a more faithful reconstruction of the original image.

To verify this assumption, we design two reconstruction tasks: image-text-image and text-image-text, shown in Figure 1. For the first reconstruction task, we evaluate the similarity between the semantics of the generated image and the input image, e.g., by computing the distance of image features extracted with a pretrained CLIP image encoder [42]. Afterward, we compare the generated text with human-annotated captions to assess the quality of the generated text [53]. Our experiments reveal that the quality of the reconstruction depends on the quality of the generated text. This leads to our first finding: the best description for an image is the description that enables the generative model to recreate the original image. Similarly, we design the reverse task where SD generates an image from a given text, and subsequently, BLIP produces a text from the generated image. We find that the best image representation for text is the one that generated the original text. We conjecture that through the reconstruction task, information on the input image is well preserved in the textual description and that meaningful description leads to a faithful recovery back to the image modality.

Based on our findings, we propose a novel finetuning framework that facilitates communication between text-to-image and image-to-text models, enabling them to talk to each other. Specifically, in our framework, a generative model not only receives training signals from human labels but also from a reconstruction loss. For a given image or text, one model first generates a representation of the input in the other modality and then the other model converts this representation back to the input modality. The reconstruction part forms a regularization loss to guide the finetuning of the first model. In this way, they acquire not only human supervision but also self-supervision that the generation should lead to a more accurate reconstruction. For example, the image captioning model should favor captions that not only match the labeled image-text pairs but also those that can lead to reliable reconstructions.

Our work is closely related to inter-agent communication. Language is a major means of exchanging information between agents. But how can we be sure that the first agent has the same understanding of what a cat or a dog is as the second agent? In this paper, we have the first agent analyze an image and produce a text describing that image. The second agent then obtains the text and simulates an image based on the text. This latter process can be thought of as an embodiment process [52]. We propose that communication is successful if the image simulated by the second agent is close to the image the first agent received as input. In essence, we test the effectiveness of language, which is the main communication venue of humans.

We conduct experiments leveraging the off-the-shelf models [42, 35, 10, 45, 46, 64], in particular recently developed large-scale pre-trained image captioning models [35, 34] and image generation models [46, 64]. Extensive experiments demonstrated the advantages of our proposed framework for various generative models, in both training-free and finetuning settings. Specifically, in the training-free paradigm, our framework significantly enhanced the caption and image generation, whereas, for finetuning, we achieved improved results for both generative models. Our main contributions are summarized as follows:

- **Framework:** To the best of our knowledge, we are the first to explore the communication of standard alone image-to-text and text-to-image generative models through human-interpretable text and image representations. Contrastively, related work unifies image and text generation implicitly through an embedding space.
- **Findings**: We find that the quality of a caption can be evaluated by assessing the image reconstruction produced by a text-to-image model. The best caption for an image is one that leads to the most accurate reconstruction of the original image. Similarly, the best image for a caption is the image that leads to the best reconstruction of the original text.
- **Improvements**: Based on our findings, we propose a unified framework to enhance both the image-to-text and text-to-image models. This involves finetuning the image-to-text model using a reconstruction loss computed by a text-to-image model as regularization, and finetuning the text-to-image model using a reconstruction loss computed by an image-to-text model. We analyzed and verified the effectiveness of our framework.

2. Related Work

Text to Image Generation Popular text-conditioned image generation models mainly include GAN [19, 29, 28], VAE [31, 64, 14, 18, 61], and recently developed diffusion models [24, 40, 46, 43, 47, 5]. Diffusion models model image generation as a Markov Chain and learn the reversed process, where a noise vector is gradually denoised into an image [24]. For text-guided image generation, generative models compute the conditional probability of generating an image given the text. DALL-E [43] and SD [46] are representatives of such diffusion-based models that are scaled to real-world complexity. Such large-scale generative models are trained on an extensive amount of image-text pairs like the LAION [49] dataset obtained from the web. The utilization of large-scale datasets allows these models to generate a vast diversity of images from intricate text inputs. The focus of this work is specifically on examining the mutual understanding of these large-scale models.

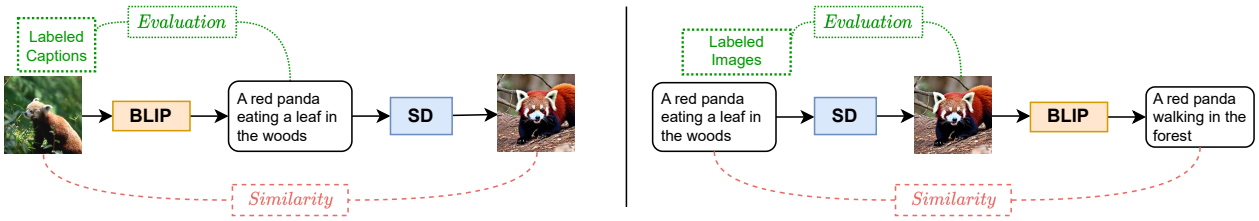

Figure 2. Illustration of our proposed inference framework. Left: a pipeline for image-text-image. The input image is fed to BLIP for caption generation and SD reconstructs the image from the generated text. The generated image is compared with the input image using a similarity function, e.g., based on image embeddings, which is utilized to evaluate the quality of the generated caption. We treat human-annotated captions as ground truth representations of the input image. Right: a pipeline for text-image-text. The reverse task for reconstructing text is demonstrated here.

Image Captioning Image captioning describes a scene using natural language [26, 56, 55, 62, 37, 60]. As one of the representatives of image captioning models, BLIP consists of an image encoder to understand the image features and a text decoder to generate text in an autoregressive manner. The image encoder uses a vision-transformer [15] backbone, which divides an image into a sequence of patches and outputs a sequence of embeddings that serve as the grounding information for text generation. Thereafter, a text decoder predicts the next token by attending to previous tokens and the encoded visual states. A recent trend in image captioning research is to develop large-scale visual language models that unify text generation with multiple image-text understanding tasks, e.g., image-based question answering and image-text retrieval [2, 13, 60]. We utilize BLIP as the image captioning method in this work.

Vision Language Representation Learning and Understanding Representation learning for vision and language handles semantic alignment between different modalities [51, 12, 13, 66, 7, 21]. A popular approach is to utilize contrastive learning on large-scale image and text pair datasets to obtain a unified representation of different representations for the same concept [42]. For unimodal image representation, self-supervised visual encoders, e.g., DINO [10], are proposed to encode visual semantics. For unimodal text representation, language models such as SBERT [45] can well extract semantics from the text. Several works have been proposed to enhance image and text generation by leveraging aligned image and text representations in the embedding space [38, 6, 58]. In contrast to their work, we aim to understand the representations learned by different image-to-text and text-to-image generative models using text or image as an interface.

3. Findings: Do BLIP and Stable Diffusion Understand Each Other?

In this section, we introduce two reconstruction tasks to test the mutual understanding of frozen BLIP and SD. We present our findings that a training-free reconstruction process can enhance the quality of a generated text or image.

3.1. Task Setup

Image-Text-Image This task aims to study the relationship between the quality of a reconstructed image and that of a caption. As shown on the left side of Figure 2, BLIP describes a given image in text, and SD generates an image from this description. Firstly, BLIP describes a source image x with different caption candidates $\{y^{(i)}\}_{i=1:N}$. When generating the captions, we use Top-p sampling [25] as the default sampling method while also examining two additional sampling strategies. Secondly, the captions are given to the SD to generate one image $\hat{x}^{(i)}$ for each caption $y^{(i)}$. We acquire N generated images for each input image. Finally, we compute the similarity between a generated image and its corresponding input image, as shown in red dashed lines in Figure 2. We utilize image embeddings obtained from pretrained encoders such as CLIP image encoder and DINO. The image encoder first encodes the input image x and a generated image $\hat{x}^{(i)}$ into embeddings space and the cosine similarity between the two embeddings is computed. We also experiment with additional image fidelity measures for a more solid evaluation.

Text-Image-Text Likewise, in the setting of *SD talking to BLIP*, shown on the right side of Figure 2, we investigate whether an improved text reconstruction corresponds to a superior image representation for a given text. We randomly sample a text from the image-text pair dataset and generate N images for each text using SD. Then BLIP generates a description for each input image using beam search [17]. Following this, we use different methods to calculate the similarity between the input and generated text, which is shown as the red dashed line in Figure 2. The similarity is computed with a text encoder where the alignment between two text embeddings extracted by the encoder is calculated. Two encoders are applied in our work. The first one is a multimodal text encoder, i.e., CLIP text encoder trained to align text and image features; The second one is a unimodal text encoder, i.e., SBERT [45] trained only on text data. Besides, traditional text similarity metrics like CIDEr [53] and WMD [33] between texts are also applied.

3.2. Evaluation Protocols

Evaluation Metric *Image Captioning:* We report standard metrics, including BLEU [41], CIDEr [53], SPICE [3], and WMD [33], as well as an embedding-based metric that computes the cosine distance between the embeddings of a candidate text and a human-labeled reference caption. For that, the CLIP text encoder and SBERT text encoder are used to obtain the caption embeddings. *Image Generation:* We use standard fidelity scores like FID [23] and Inception Score (IS) [48] to quantify the quality of the generated images. Similar to the caption evaluation, we also report embedding-based image distances between a generated and a real reference image. For that, we report the CLIP Visual Score similar to [32] and the DINO Score [10].

Implementation Details We set the number of caption candidates to N=10. The NoCaps [1] validation set and the COCO Karpathy test split [27] are used to support the evaluation. We randomly sample a caption for each image in the dataset for the input of the text-image-text task. All models are frozen in this section, without any finetuning. More implementation details can be found in Appendix A.

3.3. Findings

Insight I: a Better Caption is the One That Leads to a Better Visual Reconstruction. In the setting of image-text-image, we compute the captioning score of each generated text description, as well as the similarity score between the corresponding generated image and the input image. Then we rank the image similarities within N generated pairs for each input image and then aggregate the caption scores over the entire dataset. Figure 3 displays the correlation between the quality of the image reconstruction and the caption quality. Four scoring metrics and three similarity methods are analyzed in this experiment.

It is evident from Figure 3 that the reconstruction quality evaluated by a text-to-image model reveals the quality of the generated caption for the input image. Specifically, the better the reconstructed image, the better the caption score, independent of the similarity and the evaluation metrics. Moreover, DINO shows on-par performance with CLIP, even though DINO is trained only with image augmentations and contrastive learning on the pixel space. This rules out the possibility that CLIP might influence the comparison since it is trained to align multimodal image-text features. Furthermore, we obtain consistent results for both embedding-based captioning metrics and word frequency matching metrics.

Quantitatively, we compare our method with the baseline sampling method. For baseline, we repeat the sampling multiple times and average the final captioning metric. As shown in Table 1, our approach consistently produces better captions. The relative gain is more significant for NoCaps, especially for the out-domain subset, which contains novel objects that usually require accurate concepts to describe. The choice of the hyperparameter for the baseline sampling method is discussed in Appendix B. Additionally, we find our conclusion consistent for two different sampling strategies widely used for captioning (See Appendix C). The number of candidates N has minimal impact on the conclusion and is discussed in Appendix D.

Qualitatively, we observed plausible improvements. Our approach accurately describes the bird in Figure 5 with detailed attributes. More figures are in Appendix E.

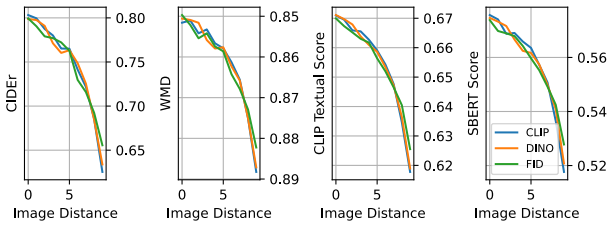

Figure 3. Evaluation of the image-text-image pipeline with three similarity and four caption metrics on the NoCaps dataset. The result suggests that regardless of the similarity or evaluation metric, better image reconstruction always leads to better captions.

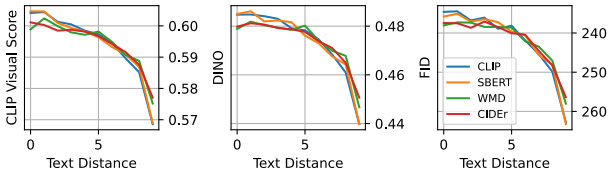

Figure 4. Evaluation of the text-image-text pipeline with four types of similarity and three types of evaluation metrics on the NoCaps dataset. The result is consistent for all types of combinations.

Insight II: a Better Image is the One That Leads to a Better Text Reconstruction. Similarly, we find that text reconstruction boosts the quality of text-to-image generation. As shown in Figure 4, when the reconstructed text is dissimilar to the original text, on average the image quality is low. In contrast, high-quality reconstruction is associated with a high-quality image. This holds true for both the fidelity and semantic metrics on image generation evaluation. For different similarity functions, we find that embedding-based similarity methods, i.e., CLIP and SBERT, perform better than similarity metrics based on word co-occurrence, i.e., CIDEr and WMD. Table 2 demonstrates the quantitative improvement of our method compared to the baseline in terms of fidelity and semantic alignment. Additionally, we present a qualitative example in Figure 5. The last image does not properly represent the input text describing a piece of cheesecake, which leads to an inaccurate caption in the next step. By comparing the generated captions with the in-

Method	Nocaps								COCO			
	In-domain		Near-domain		Out-domain		Overall		Karpathy Test			
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	B@3	B@4	CIDEr	SPICE
Baseline Sampling	75.1	12.4	72.4	11.9	78.7	11.5	74.1	11.9	32.4	21.9	90.1	19.6
Ours	**77.3**	**12.9**	**78.3**	**12.6**	**88.8**	**12.4**	**80.3**	**12.6**	**32.5**	**22.0**	**92.0**	**20.1**
Gain (%)	+2.9	+4.0	+8.1	+5.9	+12.8	+7.8	+8.4	+5.9	+0.4	+0.3	+2.1	+2.2

Table 1. Comparison of the baseline captioning model and our proposed method on Nocaps and COCO datasets. Our method outperforms the baseline sampling method on all metrics. The relative gain of our method compared to the sampling method is given in the third row. B@k: BLEU@k.

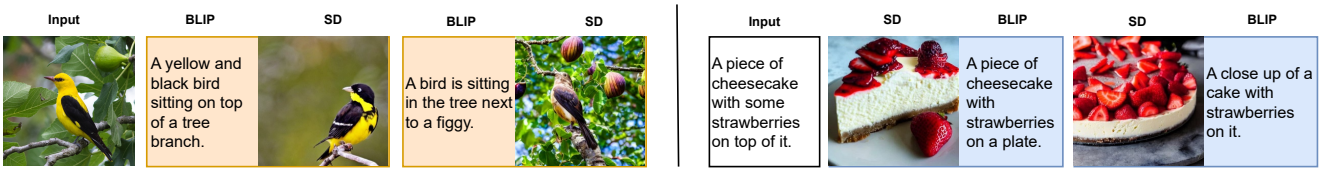

Figure 5. Qualitative examples of reconstruction tasks. The left side shows image reconstruction for the given input image of a bird. Two samples are shown with their generated images, ranked by the image similarity. The right side shows text reconstruction whereas the first sample gives a high-quality image as well as high-quality text reconstruction.

Model	NoCaps			COCO		
	CLIP↓	FID↓	IS↑	CLIP↓	FID↓	IS↑
Sampling	40.54	32.37	41.19	42.54	44.76	30.01
Ours	**33.47**	**29.59**	**45.64**	**34.74**	**42.02**	**31.62**
Gain (%)	+17.4	+8.6	+9.8	+18.3	+6.1	+5.4

Table 2. Comparison of our proposed method to SD on image generation. Our method uses BLIP to filter out generated images based on text reconstruction.

put text, we find better images that faithfully depict the text. These examples highlight the finding that an image-to-text model can be used to assess the quality of a generated image, i.e., the more similar the reconstructed text to the input text, the higher the quality of the generated image.

Additional experiments with different image captioning and text-to-image generation models in Appendix F lead to the same conclusion for the two tasks.

4. Method: Let BLIP and Stable Diffusion Talk

Based on the insights in the last section, we introduce a novel approach to finetuning the image captioning and text-to-image models by incorporating reconstructions as regularization losses. Similar to the last section, we introduce two pipelines: image-text-image and text-image-text. In the first pipeline, BLIP generates a caption for a given image, where SD is guided by this caption to reconstruct the input image. The core component is a differentiable layer that connects the output of the BLIP with the input of SD. This connection allows optimizing BLIP using the loss computed by SD in the text-to-image generation process. Likewise, in the second pipeline, we optimize the SD model using the loss obtained from comparing the generated text by BLIP with the input text. A schematic of our training framework is presented in Figure 6 and explained below.

4.1. Image-Text-Image

Text Generation Stage The left box in Figure 6 (a) illustrates the standard image captioning process using the BLIP model. BLIP takes an image $x \in \mathbb{R}^{H\times W\times 3}$ in RGB space as input and produces a sequence of tokens y_t. The prediction of each token y_t at time step t relies on tokens generated in previous steps $y_{<t}$ and the image embeddings. To accelerate the training using batch operation, a ground-truth caption y is fed to BLIP with an attention mask, ensuring that each token's prediction is causally dependent on the tokens that came before it. In this way, BLIP can generate a caption for an input image using only a single forward pass during training. Therefore, the training objective of the image captioning model is to minimize the cross-entropy loss between the ground truth text and the predicted text, defined as

$$\mathcal{L}_{TG} = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\Pi_t p_\theta(y_t|y_{<t}, x)\right], \quad (1)$$

where θ refers to the weights of BLIP and $\mathcal{D}$ refers to the dataset from which a ground-truth image-text pair (x, y) is sampled. The BLIP model was pretrained using $\mathcal{L}_{TG}$. Throughout the finetuning process, this loss is further utilized to update the weights of BLIP, preventing its predictions from deviating from the text used in pretraining.

Differentiable Connection The final softmax layer of BLIP generates a token distribution $\hat{g}_t \in \mathbb{R}^V$ at each timestep which is used to decode a discrete token y_t during image captioning. V represents the size of the vocabulary of a tokenizer. However, instead of sampling a specific y_t at each step, we store the token distributions $\hat{g} \in \mathbb{R}^{L\times V}$ for all timesteps, where L is the number of tokens in a caption. During the training phase, this can be obtained after a forward process of the input image. Subsequently, we compute the dot product of the token distributions $\hat{g}$ with the vo-

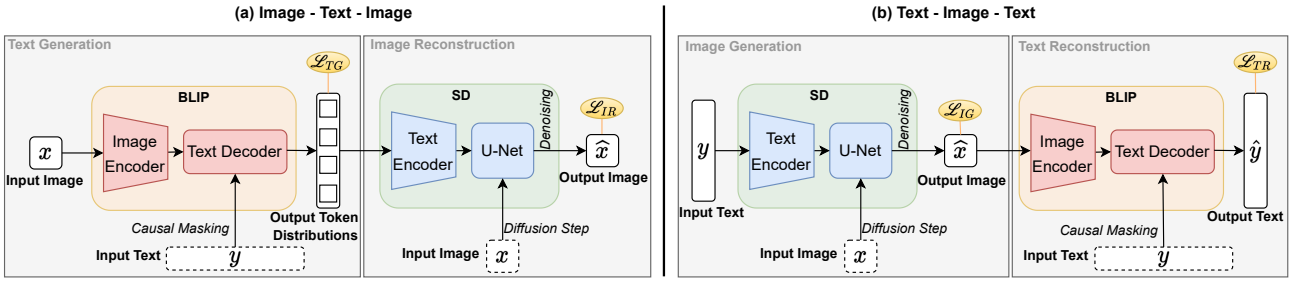

Figure 6. Illustration of our proposed finetuning framework. We introduce a reconstruction pipeline starting from an image (left) and a reconstruction pipeline starting from the text (right). Input images and text shown at the bottom (dashed boxes) are only used during training and will be dropped during inference. For illustration proposes, we omit the VAE in SD without changing the principles. Causal masking is a technique for language modeling during training. The diffusion step denotes adding noise to the input image.

cabulary embeddings $E \in \mathbb{R}^{V \times D}$ of SD's tokenizer. This produces text embeddings, which are used to guide image generation. Later stages follow the standard text-to-image generation pipeline.

In addition, to address the discrepancy between the tokenizers used by BLIP and SD, we employ a hard-coded transformation matrix to map the vocabulary of BLIP's tokenizer to that of SD (See Appendix G). For future work, we will explore discrete sampling methods for caption generation. In this work, we stick to our simple strategy as it has proven to be useful in practice.

Image Reconstruction Stage The right box of Figure 6 (a) shows the conventional procedure of training a text-guided image generative diffusion model. The model is trained on an input image x along with its textual description y. The training procedure involves a forward diffusion process wherein a clean image is progressively destructed by introducing noise in iterative steps. Then SD learns a reversed process where it predicts the noise that is necessary to reconstruct the denoised image for each step. In the context of text-to-image generation, the input text y is processed by the text encoder as conditional information to guide this reversed denoising process.

Specifically, for the diffusion process, a noise vector ϵ for timestep t is sampled from a normal distribution. A noised image x_t of the original clean image x_0 for timestep t is computed following a predefined noise scheduling process denoted as $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon$, where α_t is a predefined scalar value for noise scheduling. The U-Net [2] of SD predicts ϵ from x_t, t, and the encoding of the text input c. Formally, the training objective of SD is to minimize a mean squared error (MSE), defined as $\mathcal{L} = \mathbb{E}_{x,y\sim\mathcal{D},t\sim[1,T]}\Big[\|\hat{\epsilon}_\psi(x_t, t, c) - \epsilon\|^2\Big]$, where ψ represents the weights of SD and T is a fixed number for diffusion steps. SD uses a CLIP text encoder π to encode the text y, written as $c = \pi(y)$. In our proposed approach, the encoding is obtained using $c = \pi(\hat{g})$, which incorporates the modification introduced in the preceding subsection. In summary, the image reconstruction loss is,

$$\mathcal{L}_{IR} = \mathbb{E}_{x,y\sim\mathcal{D},t\sim[1,T]}\Big[\|\hat{\epsilon}_\psi(x, t, \pi(\hat{g})) - \epsilon\|^2\Big]. \quad (2)$$

Intuitively, if a caption is of high quality in describing the content of the input image, SD is expected to have a lower loss in reconstructing that image. During training, we uniformly sample a timestep t for each predicted caption.

In contrast to the approach outlined in Section 3.1, our novel design does not require an extra image encoder since the input image is already incorporated into the generation process of SD. This aligns with the standard training procedure, where SD reconstructs a given ground-truth image rather than generating a new image from scratch. Directly converting the approach in Section 3.1 into a training framework would be impractical as it requires a costly sampling procedure.

4.2. Text-Image-Text

Image Generation Stage This stage is equivalent to the conventional training of SD described above. Similar to Eq. 2, the loss is defined as

$$\mathcal{L}_{IG} = \mathbb{E}_{x,y\sim\mathcal{D},t\sim[1,T]}\Big[\|\hat{\epsilon}_\psi(x_t, t, \pi(y)) - \epsilon\|^2\Big]. \quad (3)$$

However, this procedure does not directly produce a clean image, which is required as input for BLIP. For that, we adopt the 1-step approximation [36, 54] technique for diffusion models. Specifically, based on the noisy image x_t and SD's predicted noise $\hat{\epsilon} = \hat{\epsilon}_\psi(x_t, t, \pi(y))$, a clean image $\hat{x}_0$ can be recovered by $\hat{x}_0 = \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\alpha_t}\hat{\epsilon})$. During training, for each sampled timestep t, we predict $\hat{x}_0$ and feed the prediction to BLIP for caption generation.

Text Reconstruction Stage The BLIP takes the predicted image $\hat{x}_0$ as input and produces a caption. Following the

[2] In fact, SD utilizes a variational autoencoder to compress the image into a latent space and subsequently learns the diffusion process in that space. For more details, readers are referred to [46].

Method	Nocaps									COCO			
	I-C	N-C	O-C	B@1	B@2	B@3	B@4	CIDEr	SPICE	B@3	B@4	CIDEr	SPICE
BLIP ViT-B [35]	111.8	108.6	111.5	83.6	68.2	50.6	32.0	109.7	14.7	50.5	39.7	133.3	23.8
Ours ViT-B, SD	**114.9**	**110.8**	**112.9**	**84.6**	**69.4**	**51.9**	**32.9**	**111.8**	**14.9**	**51.4**	**40.1**	**134.6**	**24.0**
BLIP ViT-L [35]	114.9	112.1	115.3	84.2	69.3	51.7	33.1	113.2	14.8	51.4	40.4	136.7	**24.3**
Ours ViT-L, SD	**116.1**	**113.0**	**115.6**	**84.7**	**70.0**	**52.5**	**33.6**	**114.0**	**14.9**	**52.0**	**40.9**	**137.3**	24.1
BLIP-2 [34]	122.7	118.0	123.8	86.8	73.2	56.2	37.0	119.9	**15.4**	55.5	43.7	145.8	**25.2**
Ours BLIP-2, SD	**123.8**	**119.3**	**124.2**	**86.9**	**73.5**	**56.5**	**37.1**	**121.0**	**15.4**	**56.0**	**44.1**	**146.4**	**25.2**

Table 3. Evaluation results of image captioning. We conduct experiments with different image captioning models (BLIP ViT-B, BLIP ViT-L, BLIP-2) with an image generation model (SD). I-C/N-C/O-C: In-/Near-/Out-domain CIDEr. The remaining metrics are on the entire set. B@k: BLEU@k.

same procedure in the text generation stage, the text reconstruction loss is defined as

$$\mathcal{L}_{TR} = \mathbb{E}_{x,y\sim\mathcal{D}}\big[\Pi_t p_\theta(y_t|y_{<t}, \hat{x}_0)\big]. \tag{4}$$

The major difference is that $\mathcal{L}_{TR}$ is conditioned on images with gradients originating from SD, which allows optimizing SD's parameters by BLIP's loss.

4.3. Full Training Objective

We simplify the training pipelines by adding the individual losses into a single training loss, and then optimizing both models on this summed loss. The parameters of both models are updated at each iteration, enabling a joint improvement of both models.

$$\mathcal{L}(\theta, \psi) = \mathcal{L}_{TG} + \mathcal{L}_{IR} + \mathcal{L}_{IG} + \mathcal{L}_{TR}. \tag{5}$$

See Appendix H for a discussion on our loss function, including its connection to CycleGAN [65], and the pseudo-code for the training framework.

5. Experimental Setup

Dataset and Evaluation *Training dataset:* We finetune both BLIP and SD on the COCO Karpathy train split [27] of 113k images, each associated with five captions. *Image Captioning:* Following [35], the image captioning is evaluated on the COCO test set and NoCaps validation set, utilizing metrics described in Section 3.2. Unlike the COCO dataset, which contains images with common object categories, the NoCaps dataset includes images in the wild, making it a challenging benchmark for zero-shot evaluation for image captioning. *Image Generation:* For image generation, we report the CLIP image distance, defined in Section 3.2, on the COCO test set and the NoCaps validation set. Likewise, the NoCaps dataset is used as a zero-shot evaluation benchmark. As each image is associated with multiple captions, we randomly sample a caption for each image to construct these two test sets.

Baselines and Models *Image Captioning:* BLIP ViT-B denotes the BLIP model with ViT-B/16 as the visual backbone. BLIP ViT-L uses a larger variant of ViT, i.e., ViT-L/16. We use the bootstrapped and finetuned version since they are optimized to produce the best baselines for captioning. We further explore the BLIP-2 ViT-G $\text{OPT}_{2.7\text{B}}$ which is among the SOTA captioning methods. We do not further finetune their models as they have already been finetuned on the same dataset using text generation loss. Thus we use performance metrics reported from their paper [35, 34]. For the metrics that are not available, we run the evaluation with their published code and weights to get the results. *Image Generation:* Since, as of this writing, we have not found any other publicly available text-to-image generation models with on-par performance, we utilize SD as the baseline model. We use the weights of sd-v1-4.ckpt for SD. The output image size is 512 except for BLIP-2, where we downscaled the image size to 384 due to hardware constraints. Since SD is not trained on the COCO dataset, we finetune SD using MSE loss to serve as a baseline method.

For our finetuning framework, we conducted experiments with the combinations of the three above-mentioned image captioning models and the text-to-image model. For the ablation study, we use BLIP ViT-B as the default setting.

Training Details To improve the efficiency, we finetune a subset of weights of SD and BLIP following heuristics [32] and hyperparameter search. For BLIP ViT-B and ViT-L, we finetune the query projection weights in the cross-attention layer of the text decoder and freeze the other components. For BLIP-2, the query tokens are adapted. For SD, we finetune the query projection weights in the cross-attention layer. Unless specified, we use a learning rate of 1e-4 and batch size of 8 and finetune the framework for 5 epochs. The experiments are conducted on an A10 GPU with 24GB of memory. The finetuning is efficient, requiring only a forward pass for both BLIP and SD per input. Thus, the additional computations introduced by the reconstruction process are relatively minor, amounting to approximately 1.4 times the original cost.

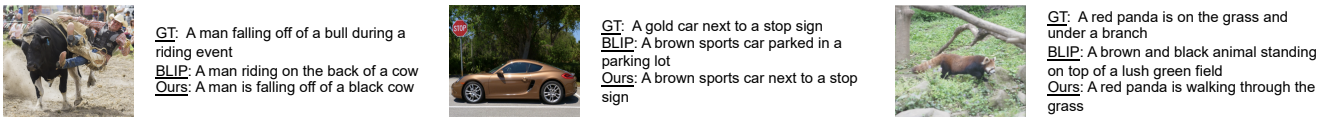

Figure 7. Qualitative evaluation of image captioning. From top to down, we show a ground-truth (GT) caption, a generated caption from BLIP ViT-B, and a generated caption from our finetuned model of ViT-B.

Figure 8. Qualitative evaluation of image generation. The left side shows a textual input and a reference image. The upper images on the right side are generated by the SD baseline. The lower images are generated from our model.

Model	NoCaps CLIP ↓	NoCaps FID ↓	COCO CLIP ↓	COCO FID ↓
SD [46]	0.4039	**21.19**	0.4248	27.08
SD MSE	0.4077	24.32	**0.4011**	25.66
Ours ViT-B, SD	**0.3978**	21.96	0.4068	24.68
Ours ViT-L, SD	0.4031	22.92	0.4071	**24.21**

Table 4. Evaluation results of image generation. We compare baseline methods, SD and MSE loss, with our image generation models that are finetuned with two image captioning models.

6. Evaluation Results

Here we present a comprehensive evaluation of our framework by comparing it to baselines and showing qualitative results. In addition, we discuss and analyze our training objective.

Improvement in Image Captioning Table 3 shows the performance of different models on the image captioning task. We report a broader set of metrics on the entire dataset for a more detailed evaluation. As shown in the table, our finetuned model demonstrated improved performance across most metrics. Our approach is effective across different sizes (ViT-B vs ViT-L) and architectures (BLIP vs BLIP-2) of image captioning models. Note that our model outperforms ViT-B with a 2% improvement in CIDEr, which is a larger gain compared to BLIP's 0.7% improvement obtained through bootstrapping on a dataset of 129M images.

Improvement in Image Generation Similarly, we compare the performance of text-to-image generation models in Table 4. We observe that SD finetuned with our reconstruction loss can produce more semantically aligned images on the NoCaps dataset that contains diverse and real-world scenarios, despite slightly reduced fidelity. Finetuning SD solely with MSE loss on human-labeled image-text pairs yields improvements in the COCO test set, which has the same data distribution as the training data. However, this approach leads to worse performance in image fidelity and degraded alignment on the NoCaps dataset. In general, utilizing reconstruction loss leads to improvements over human supervision. Our proposed framework has demonstrated significant potential to improve image generation.

Qualitative Results Our method generates a more faithful description of a scene in terms of relationships, context, and fine-grained concepts. For example, BLIP correctly recognizes the objects in the left-most image in Figure 7, but it fails to predict the relationship between them. In contrast, our model generates a correct description, likely driven by the reconstruction loss that the visual perception of a man riding a cow is different from that of a man falling off a cow. In addition, Figure 8 shows an example of generated images using our model and SD baseline method. The baseline may neglect certain objects like the car, whereas our method reflects the text prompt. More figures are in Appendix J.

Captioning			Image Generation		
Model	CIDEr	SPICE	Model	CLIP ↓	FID ↓
Ours	**111.8**	**14.9**	Ours	**0.3978**	**21.96**
w/o $\mathcal{L}_{TG}$	102.3	14.0	w/o $\mathcal{L}_{IG}$	0.4241	26.56
w/o $\mathcal{L}_{IR}$	109.7	14.7	w/o $\mathcal{L}_{TR}$	0.4077	24.32

Table 5. Analysis of the loss function. We conduct experiments with either only the regularization loss or only the supervision of image-text pairs.

Analysis of Loss Function We investigate the effectiveness of our proposed loss function, especially examining if the reconstruction loss alone can still lead to the improvement exhibited above. The first row in Table 5 shows the results of our final models trained on the combined pipelines using Eq. 5, whereas for the second and third rows, we separately train the two pipelines and ablate the loss terms. The second row on the left side of the table refers to the BLIP model trained solely on $\mathcal{L}_{IR}$ using the framework in Figure 6 (a). As expected, the model struggles to provide competitive results. This shows the benefit of human supervision, which prevents the model from overfitting the reconstruction objective. For completeness, we also show the performance of BLIP with only the image captioning objective in the third row of Table 5. As our training framework involves different loss terms and finetuning strategies, a more detailed analysis of those aspects can be found in Appendix I.

7. Discussions and Limitations

Besides our findings and improvements, this work also opens up avenues for potential future directions. 1) The novel design of backpropagating through SD allows image generation to be used as a downstream task, wherein the knowledge in diffusion models is effectively transferred. 2) The findings in our work inspire the development of label-free evaluation metrics for image captioning.

The primary limitation of this work is the restricted scope of improvement. In our work, the degree of improvement in BLIP and SD depends on their initial capacities. As a result, challenges may remain for complex images or intricate descriptions which are not well covered in the data distribution for pretraining. Besides, we treat the transformations from image to text and text to image as a black box, lacking a deeper understanding of the alignment between layers within models in the generation processes. We leave further exploration with explanation methods in future work [50, 4, 22]. Another under-explored perspective is the robustness of our finetuning paradigm. Concretely, it is not clear how the fine-tuned models perform under out-of-distribution images and texts [20, 11]. Additionally, our work also inherits the known limitations of large-scale generative models [2, 63], bringing concerns about possible biases or harmful content generation.

8. Conclusion

In this work, we investigated mutual understanding between multimodal image-to-text models and text-to-image models through image-text-image and text-image-text reconstruction tasks. We found that the reconstruction quality of text-to-image and image-to-text models can be utilized to evaluate the quality of the text or image generation. Specifically, the best textual description for an image is one that leads to a better reconstruction of the input image. The best image representation of text input is one that leads to a better recovery of the original text. Leveraging these findings, we proposed a novel framework for finetuning the image captioning and image generation models. We demonstrated enhanced performance of our models on both tasks. Our work advocates further exploring multimodal communication between text-to-image and image-to-text models. Finally, our work demonstrated the value of symbolic sentences to convey information: Image content can effectively be compressed into a sentence, and a sentence can be reconstructed as an image. This latter step can be considered a form of grounded cognition or embodiment.

References

[1] 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/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 4, 13

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 3, 9

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 4

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 9

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[6] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023. 3

[7] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 3

[8] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021. 1

[9] Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. *arXiv preprint arXiv:1811.02549*, 2018. 13

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3, 4

[11] Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *arXiv preprint arXiv:2306.02080*, 2023. 9

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 3

[13] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 3

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao,

Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[16] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. 13

[17] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *ACL 2017*, page 56, 2017. 3

[18] 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. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[20] Jindong Gu, Ahmad Beirami, Xuezhi Wang, Alex Beutel, Philip Torr, and Yao Qin. Towards robust prompts on vision-language models. *arXiv preprint arXiv:2304.08479*, 2023. 9

[21] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 1, 3

[22] Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 119–134. Springer, 2019. 9

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[25] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 3

[26] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1, 3

[27] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4, 7, 13

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[30] Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, and Kyunghoon Bae. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16526–16536, 2022. 1

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 2

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 4, 7

[33] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France, 07–09 Jul 2015. PMLR. 3, 4

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 7, 14

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2, 7, 13, 14

[36] Wei Li, Xue Xu, Xinyan Xiao, Jiachen Liu, Hu Yang, Guohao Li, Zhanpeng Wang, Zhifan Feng, Qiaoqiao She, Yajuan Lyu, et al. Upainting: Unified text-to-image diffusion generation with cross-modal guidance. *arXiv preprint arXiv:2210.16031*, 2022. 6

[37] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 1, 3

[38] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 1, 3

[39] Moin Nadeem, Tianxing He, Kyunghyun Cho, and James Glass. A systematic characterization of sampling algorithms for open-ended language generation. *arXiv preprint arXiv:2009.07243*, 2020. 13

[40] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[41] 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. 4

[42] 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. 1, 2, 3

[43] 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. 1, 2

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 2, 3

[46] 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. 1, 2, 6, 8, 13, 14

[47] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[48] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 4

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[50] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 9

[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 3

[52] Volker Tresp, Sahand Sharifzadeh, Hang Li, Dario Konopatzki, and Yunpu Ma. The tensor brain: A unified theory of perception, memory, and semantic decoding. *Neural Computation*, 35(2):156–227, 2023. 2

[53] 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. 2, 3, 4

[54] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. *arXiv preprint arXiv:2303.13703*, 2023. 6

[55] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 1, 3

[56] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 3

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1

[58] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 3

[59] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022. 1

[60] 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. 1, 3

[61] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[62] 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. 1, 3

[63] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained trans-

former language models. *arXiv preprint arXiv:2205.01068*, 2022. 9

[64] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021. 2, 14

[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 7

[66] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. 3

[67] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 1