LEVERAGING UNPAIRED DATA FOR VISION-LANGUAGE GENERATIVE MODELS VIA CYCLE CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Current vision-language generative models rely on expansive corpora of *paired* image-text data to attain optimal performance and generalization capabilities. However, automatically collecting such data (e.g. via large-scale web scraping) leads to low quality and poor image-text correlation, while human annotation is more accurate but requires significant manual effort and expense. We introduce **ITIT** (**I**n**T**egrating **I**mage **T**ext): an innovative training paradigm grounded in the concept of cycle consistency which allows vision-language training on *unpaired* image and text data. ITIT is comprised of a joint image-text encoder with disjoint image and text decoders that enable bidirectional image-to-text and text-to-image generation in a single framework. During training, ITIT leverages a small set of paired image-text data to ensure its output matches the input reasonably well in both directions. Simultaneously, the model is also trained on much larger datasets containing only images or texts. This is achieved by enforcing cycle consistency between the original unpaired samples and the cycle-generated counterparts. For instance, it generates a caption for a given input image and then uses the caption to create an output image, and enforces similarity between the input and output images. Our experiments show that ITIT with unpaired datasets exhibits similar scaling behavior as using high-quality paired data. We demonstrate image generation and captioning performance on par with state-of-the-art text-to-image and image-to-text models with orders of magnitude fewer (only 3M) paired image-text data.

1 INTRODUCTION

Image-text multimodal training has gained remarkable attention in recent years. Models for text-to-image generation show the impressive ability to synthesize realistic images from textual prompts (Rombach et al., 2022; Chang et al., 2023; Saharia et al., 2022; Yu et al., 2022b; Ramesh et al., 2021). Similarly, image-to-text models have demonstrated advanced image comprehension capabilities by providing precise descriptions of input images (Chen et al., 2023; Wang et al., 2022a; Li et al., 2022; 2023a; Alayrac et al., 2022; Wang et al., 2022b). Training these models to exceptional performance demands datasets comprising hundreds of millions to billions of *paired* image-text samples (Schuhmann et al., 2022). The collection of very large paired datasets comes at a considerable cost as well as concerns of low quality (Jia et al., 2021). On the other hand, diverse and vast unimodal images or texts datasets remain unused in current generative vision-language training (Raffel et al., 2020; Sun et al., 2017; Zhai et al., 2022). This raises a natural question: can we leverage *unpaired* image and text data to facilitate generative vision-language training?

The major problem with using unpaired data during vision-language training is the lack of supervision. To overcome this problem, we introduce ITIT, a novel training paradigm that uses *cycle consistency* losses between cycle-generated images/texts and their corresponding original inputs to provide supervision for image-only and text-only data (Figure 1). ITIT utilizes a small set of paired image-text data to achieve reasonable text-to-image and image-to-text generation performance. Simultaneously, for unpaired image (text) data, ITIT generates corresponding text (image) counterparts and employs them as inputs to reconstruct the input image (text): this corresponds to a full cycle loss. We consider two kinds of full cycles: T2I2T (starting with an unpaired text sample); and I2T2I (starting with an unpaired image sample). These two types of cycles enable us to leverage both unpaired image and text data to provide informative supervision signals for training.

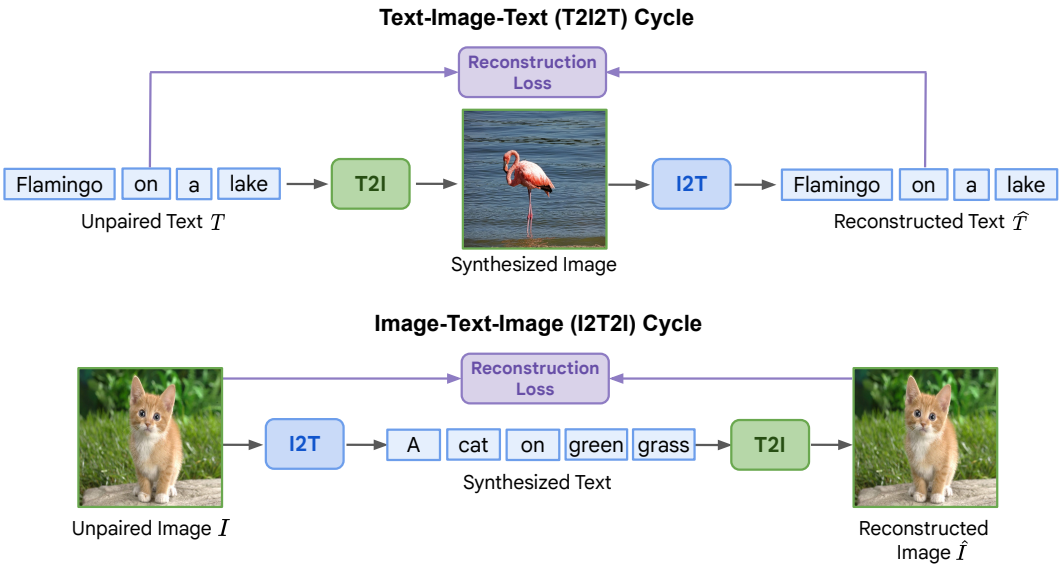

Figure 1: Overview of ITIT. For unpaired data, ITIT first generates the image/text counterpart, and then uses these generated counterparts to reconstruct the original text or image.

To enable cycle training, we first unify image-to-text (I2T) and text-to-image (T2I) generation in the same framework, with a bi-directional image-text encoder and disjoint image and text decoders. We tokenize images into discrete visual tokens (Van Den Oord et al., 2017) and combine them with text embeddings from a pre-trained T5 model (Raffel et al., 2020) as input to the joint image-text encoder. For I2T generation, we employ an autoregressive text decoder (Wang et al., 2022a), while for T2I generation we use a non-autoregressive parallel image decoder (Chang et al., 2023), which is an order of magnitude faster than autoregressive image decoders such as Yu et al. (2022b).

A technical challenge of ITIT is that, state-of-the-art text-to-image and image-to-text generation processes typically involve multiple forward steps of the model (Esser et al., 2021; Chang et al., 2023; Rombach et al., 2022; Wang et al., 2022a). Back-propagating gradient through all these forward steps brings significant memory and computation overheads. To solve this problem, for T2I2T cycle, we first generate the image with parallel decoding. We then back-propagate the gradient through one step of the parallel decoding process. For I2T2I cycle, we first generate the text autoregressively with multiple steps. Then we forward the text decoder once with the generated text as input, and back-propagate the gradient only to this forward step. This significantly reduces the computational overhead of the cycle training, making it feasible to apply in large model settings.

We evaluate the performance of ITIT on standard image-to-text and text-to-image generation benchmarks and demonstrate that, by leveraging unpaired data and cycle consistency, ITIT attains performance levels similar to a non-cycle baseline. However, ITIT uses up to 2 orders of magnitude lower paired data. Furthermore, ITIT scales similarly with unpaired data as the baseline does with equivalent amounts of paired data, while being much more robust to low data quality. We also compare ITIT with state-of-the-art methods and show that we can achieve comparable performance on common text-to-image and image-to-text benchmarks with substantially lesser paired data. Our contributions are summarized as follows:

- We introduce a framework that unifies text-to-image and image-to-text generation, and propose ITIT, a novel technique that enforces consistency between cycle-generated images/text and their corresponding originals. This approach allows the training of image-to-text and text-to-image models using unpaired image and text data.
- We comprehensively evaluate the proposed ITIT framework and the image-text cycle consistency method, and demonstrate that they significantly enhance model performance.
- We show that ITIT can achieve performance on par with state-of-the-art methods on common text-to-image and image-to-text benchmarks with much lesser ($\sim$100x) paired data. When scaling up training data to improve model efficacy, we show that we can add only unpaired examples using our framework and achieve similar performance as scaled-up paired data, without the downsides of significant manual effort and poor pairing quality.

2 LITERATURE REVIEW

Image-to-Text Generation. Various works explore autonomously generating textual descriptions from input images, either training the network with generative loss alone (Wang et al., 2022b; Alayrac et al., 2022; Chen et al., 2023; Li et al., 2022; 2023a), or combining it with contrastive learning (Yu et al., 2022a). GIT (Wang et al., 2022a) trains a model comprising an image encoder and an auto-regressive text decoder using a language modeling loss, the image encoder pre-trained with contrastive loss (Radford et al., 2021). In our work, we adopt a similar framework to GIT for our Image-to-Text (I2T) framework, but we initialize our image encoder from scratch.

Text-to-Image Generation. Recent works focus on two primary paradigms: diffusion-based models (Rombach et al. (2022); Dhariwal & Nichol (2021); Nichol et al. (2021); Saharia et al. (2022); Ramesh et al. (2022); Ruiz et al. (2023)); and token-based methods. Token-based strategies transform raw images into image tokens, and predict these tokens either in an autoregressive manner (Esser et al., 2021; Ramesh et al., 2021; Gafni et al., 2022; Yu et al., 2021; Ding et al., 2021; Yu et al., 2022b) or in parallel (Chang et al., 2022; Li et al., 2023b; Chang et al., 2023). Muse (Chang et al., 2023) demonstrates that token-based strategies with parallel decoding can be considerably faster than diffusion-based or autoregressive generative models. Since this speed advantage facilitates text-to-image synthesis during training, we adopt this strategy in our T2I framework.

Unifying Image and Text Generation. COBIT (You et al. (2023)) achieves this by employing distinct image and text unicoders, coupled with a unified cross-modal decoder. Additionally, CM3 (Aghajanyan et al. (2022)) and CM3Leon (Yu et al. (2023)) harness causally masked generative models trained on extensive multi-modal document datasets, and enable the synthesis of both text and images. However, all these works still heavily rely on large-scale *paired* image-text datasets.

Leveraging Unpaired Data in Generative Vision-Language Training. Early works have tried to use unpaired image and text to train image captioning model in an unsupervised way (Feng et al., 2019). However, the performance is relatively poor. Recent efforts in incorporating unpaired data into generative vision-language training primarily focus on pre-trained image and text encoders (Esser et al., 2021; Roberts et al., 2019). However, these applications are limited to pre-training and do not encompass the entire generative vision-language training procedure, thus providing only incremental improvements. In some cases, researchers have explored the use of text-only data to improve text decoders (Wang et al. (2022b)), utilizing text-to-text training. However, this only enhances the text decoder and not the image encoder, resulting again in constrained improvements.

Cycle-consistency. The concept of cycle consistency has previously been used to provide regularization and/or compensate for a lack of annotated data. Zach et al. (2010); Zhou et al. (2016); Godard et al. (2016); Zhu et al. (2017); Messikommer et al. (2022) explore it for computer vision applications such as learning dense correspondence, event detection, depth estimation, and image-to-image translation. Wang et al. (2023) and Goel et al. (2022) introduce regularization on image-text consistency during CLIP pre-training, aiming to learn a better visual representation. Most related to our work is Gorti & Ma (2018), which uses text-image-text cycle consistency to perform text-to-image translation, but the performance is poor. Moreover, none of the previous works has explored the potential of cycle consistency in generative vision-language training using unpaired data.

Our novel approach diverges from preceding vision-language models that heavily rely on either a large corpus of paired image-text data, or fine-tuning methods that target only text or image encoder/decoders separately. For the first time, our method facilitates the utilization of unpaired image and text data during generative vision-language training. This innovation significantly reduces the dependency on paired image-text samples during the training process, which empowers the expansion of generative vision-language training to nearly boundless text-only and image-only datasets.

3 METHOD

ITIT is the first framework that enables generative vision-language training on unpaired image-only and text-only data. It uses a simple yet effective architecture: a unified image-text encoder and two separate image and text decoders. This design seamlessly enables text-to-image and image-to-text generation in the same framework, which paves the way for text-image-text (T2I2T) and image-text-

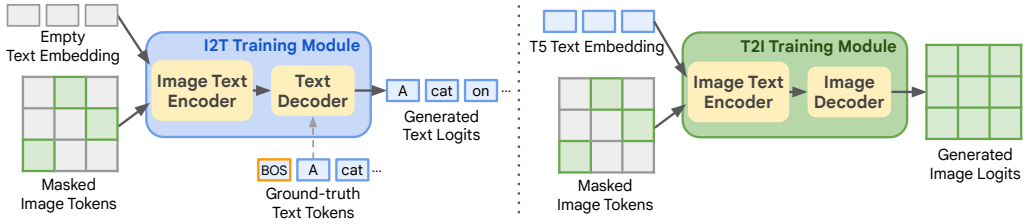

Figure 2: I2T (left) and T2I (right) training pipelines for *paired* image and text data.

image (I2T2I) cyclic losses. Below, we describe each component of our ITIT architecture and the cycle-consistency training paradigm in detail.

3.1 Unified Image-Text Generation Framework

Architecture. We first obtain text embedding $T = [t_l]_{l=1}^{L}$ from the output of a T5 encoder (Roberts et al., 2019) on the raw text. Similarly, raw images are passed through a pre-trained VQ-tokenizer (Esser et al., 2021) to output image tokens $I = [i_k]_{k=1}^{K}$. L and K are the token sequence lengths for text and image, respectively. The image tokens I are then embedded with an embedding layer and concatenated with the T5 text features T as input to the image-text encoder. Modality-specific decoders then operate on the encoded image-text features to generate either text or image tokens. The text decoder is autoregressive (Wang et al., 2022a), while the image decoder is parallel (Chang et al., 2023). Both encoder and decoders are based on Transformer (Vaswani et al., 2017) layers. A detailed description of the model architecture is included in Appendix B.

Image-to-Text (I2T) Training. As shown in Figure 2, we input masked image tokens along with empty text embedding to the image-text encoder. Masking is used to save computation, similar to MAE (He et al., 2022). We then use the features generated by the image-text encoder, as well as the ground-truth text tokens prepended with `[BOS]` (begin-of-sentence) token as the input to our text decoder. We use an auto-regressive language modeling (LM) loss to train the encoder and decoder:

$$\mathcal{L}_{I2T} = -\mathbb{E}_{(I,T)\in\mathcal{D}}\Big[\sum_{l=1}^{L}\log p(t_l|I_M, t_0, \cdots, t_{l-1})\Big], \tag{1}$$

which is a CE loss with label smoothing 0.1. Here, t_0 is set to be the `[BOS]` token. I_M are the (subset of) *unmasked* tokens in I and $p(i_k|I_M, T)$ is the probability predicted by the encoder-decoder network (the 'logits' layer), $\mathcal{D}$ is the distribution of paired image-text data. Note that the text decoder employs causal attention similar to GIT (Wang et al. (2022a)): each text token only depends on the preceding text tokens and all image features.

Text-to-Image (T2I) Training. As shown in Figure 2, right panel, we use masked image modeling for image generation, where the training objective is to reconstruct masked image tokens conditioned on the unmasked image tokens and the paired text features. We denote the binary mask determining which image tokens are masked by $M = [m_k]_{k=1}^{K}$. We use a cross-entropy loss between the ground-truth one-hot image tokens and the output of the image decoder. Specifically,

$$\mathcal{L}_{T2I} = -\mathbb{E}_{(I,T)\in\mathcal{D}}\Big[\sum_{\forall k:m_k=1}\log p(i_k|I_M, T)\Big], \tag{2}$$

Inference. We follow GIT (Wang et al., 2022a) for image-to-text inference and Muse (Chang et al., 2023) for text-to-image inference. More details are included in Appendix B.

3.2 Training with Cycle Consistency

Our cycle consistency training paradigm allows training with image-only and text-only data. The key idea is to first synthesize the corresponding text/image from the image-only or text-only data, and then use the synthesized data as input to reconstruct the original image/text. This allows us to apply cycle consistency supervision on image-only and text-only data.

Text-Image-Text (T2I2T) Cycle. Our T2I2T training pipeline is shown in Figure 3, top panel. At each training iteration, we first synthesize pseudo paired image tokens I' for input text $T = [t_l]_{l=1}^{L}$ using our T2I inference pipeline. We then apply random mask M to I', perform reconstruction

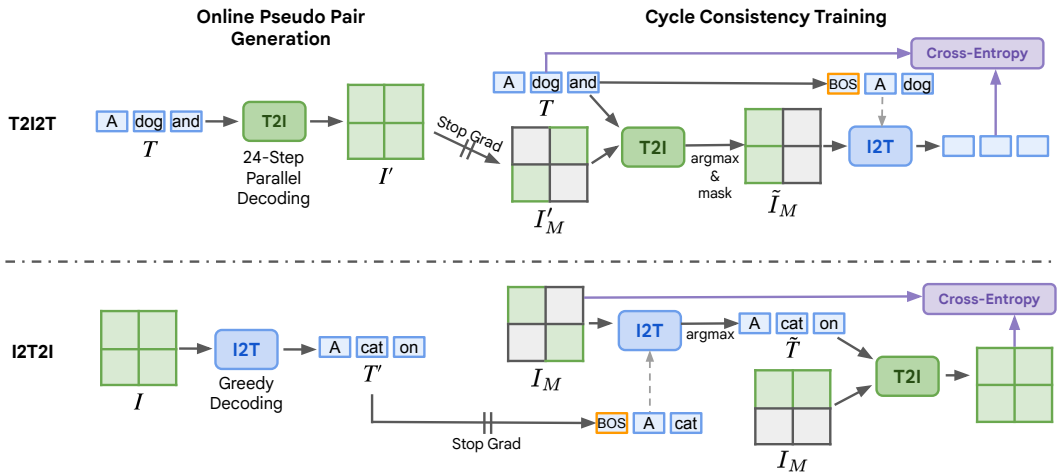

Figure 3: Text-image-text (top) and image-text-image (bottom) cycle training pipelines for *unpaired* image and text data. We use pseudo-generated image and text to enable the cycle consistency. Image token masks M are always randomly chosen. The dashed line denotes causal attention. Text tokens prepended with `[BOS]` token are used for auto-regressive language modeling loss.

on I'_M with the text T using the T2I pipeline, and obtain the reconstructed synthesized image $\tilde{I}$. This two-step process allows us to avoid the excessive memory requirements of back-propagating gradients through all 24 steps of parallel decoding, while still training the T2I module. Finally, we randomly mask $\tilde{I}$ and use $\tilde{I}_M$ to generate text using the I2T pipeline. The objective of our cycle paradigm is to enforce consistency between this generated text and the original text. Therefore, the T2I2T cycle-consistency loss can be formulated as follows:

$$\mathcal{L}_{T2I2T} = -\mathbb{E}_{T \in \mathcal{D}_{text}} \Big[\sum_{l=1}^{L} \log p(t_l | \tilde{I}_M, t_0, \cdots, t_{l-1}) \Big], \tag{3}$$

This is very similar to the I2T loss in Equation (1), except that $\tilde{I}$ is synthesized from T instead of being drawn from the image-text joint distribution.

Image-Text-Image (I2T2I) Consistency. Our I2T2I training pipeline is shown in Figure 3, bottom panel. Similar to the T2I2T pipeline, we first synthesize pseudo paired text tokens T' for input image tokens I using our I2T inference pipeline. We then use the I2T training pipeline to predict $\tilde{t_l}$ from $t'_0, \cdots, t'_{l-1}$ and I_M. As before, this avoids the excessive memory requirements of back-propagating gradients through the auto-regressive greedy decoding. We then mask I, and pass it through the T2I pipeline with the predicted $\tilde{T}$ to reconstruct the masked image tokens. Again, the loss enforces consistency between the reconstructed and the original image tokens using cross-entropy:

$$\mathcal{L}_{I2T2I} = -\mathbb{E}_{I \in \mathcal{D}_{image}} \Big[\sum_{\forall k : m_k = 1} \log p(i_k | I_M, \tilde{T}) \Big], \tag{4}$$

Gradient Estimation. One challenge in our cycle training is that $\tilde{i_k} = \arg\max(p(i_k | I'_M, T)$ and $\tilde{t_l} = \arg\max p(t_l | I_M, t'_0, \cdots, t'_{l-1})$, which are not differentiable. To solve this, we use a straight-through estimation on the predicted logits to approximate the gradient. Specifically, we directly copy the gradient on the one-hot prediction to the predicted logits after softmax. We show in section 4.4 that this helps improve both text-to-image and image-to-text performance.

4 Results

4.1 Experiment Setup

Datasets. We use three datasets in our experiments: CC3M (Sharma et al., 2018), WebLI (Chen et al., 2023), and Shutterstock (Shutterstock, 2023). CC3M contains 3.3 million high-quality image-text pairs. WebLI (Web Language Image) contains 111 million images where the image-text pairing quality is much lower than CC3M. Thus, WebLI is significantly noisier and, as we show, leads to

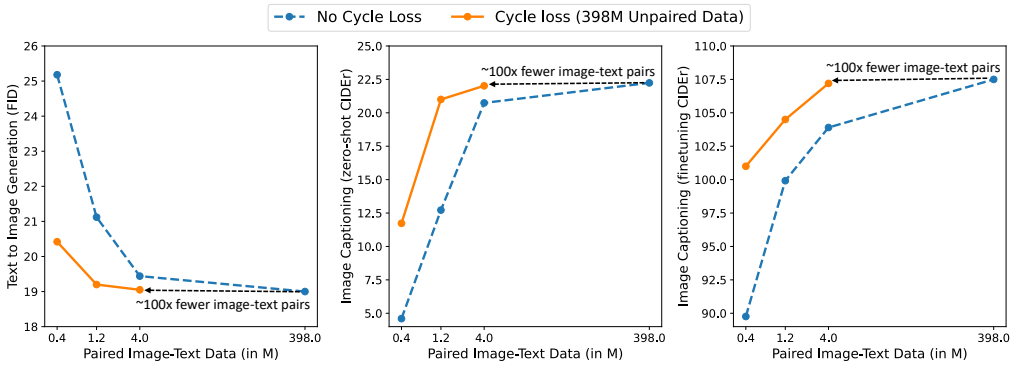

Figure 4: How ITIT-H's performance scales with additional paired Shutterstock data. The baseline (T2I+I2T) is trained with paired samples only. ITIT is trained with the same number of paired samples, as well as 398M unpaired samples (the full Shutterstock dataset) using cycle loss.

worse performance for I2T. Shutterstock contains 398 million images labeled by human annotators, which incurs significant expense and effort. More dataset details are included in Appendix C.

We use CC3M as our paired dataset, 50% of WebLI images as our unpaired image dataset, and the other 50% of WebLI texts as our unpaired text dataset for most of our experiments (Section 4.3 and Section 4.4). This 50%-50% split ensures that corresponding image-text pairs are not present in our unpaired image and text splits. We use the Shutterstock dataset in Section 4.2, where we analyze how ITIT scales w.r.t. different number of paired and unpaired data samples.

Training. We set the input image resolution as 256x256 to be consistent with previous literature. After passing through the VQGAN tokenizer, the image token sequence length is 16x16 (256 tokens). The raw text (maximum length of 64) is tokenized by SentencePiece tokenization (SentencePiece, 2023), and embedded using a pre-trained T5 encoder. These embeddings are then concatenated with the image token embeddings as the input to our image-text encoder.

We experiment with ViT-B, ViT-L, and ViT-H size Transformers (Dosovitskiy et al. (2021)) for our image-text encoder. We combine the losses in Equations 1 through 4 with equal weight for training. For results in Section 4.3, we use Adafactor (Shazeer & Stern, 2018) to train the model for 1.5M steps with a batch size of 2048 (1024 for image-text pairs, 512 for unpaired images, and 512 for unpaired texts). We use a cosine learning rate schedule with 5K steps warmup and maximum learning rate 1×10^{-4}. For other experiments, we use the exact same training paradigm except that we train the models for 500K steps. More details are included in Appendix B.

Evaluation. We follow the commonly used MS-COCO benchmark and evaluation protocols. For image-captioning, we evaluate both the zero-shot and fine-tuning performance of ITIT on the COCO Karpathy split (Karpathy & Fei-Fei, 2015) and report the CIDEr score (Vedantam et al., 2015). For text-to-image generation, we evaluate ITIT on 30K image-text pairs randomly selected from the COCO Captions training set and report the Frechet Inception Distance (FID) score (Heusel et al., 2017). CIDEr is the higher the better, and FID is the lower the better.

4.2 Scale with Data

In this section, we comprehensively evaluate ITIT's performance with different amounts of paired and unpaired data on Shutterstock dataset (Shutterstock, 2023) consisting of 398M image-text pairs.

Figure 4 analyses how ITIT's performance scales with paired data. We train a baseline with only paired data, with the sum of the losses in Equation (1) and Equation (2). ITIT is trained with the same paired data as the baseline, and the entire set of 398M images and text present in Shutterstock as unpaired data. More paired data helps both settings, but training with unpaired data significantly improves ITIT's performance over the baseline on both image captioning and text-to-image generation. Remarkably, with only 4M paired data and 398M unpaired data, ITIT achieves *a similar performance as training with 398M paired data*. Note that ITIT does not use any samples not present in the baseline trained with 398M paired data, as all of the samples are from Shutterstock. Therefore ITIT can perform similarly as a baseline with 100x fewer image-text pairs, significantly reducing the effort and expense for the training of generative vision-language training.

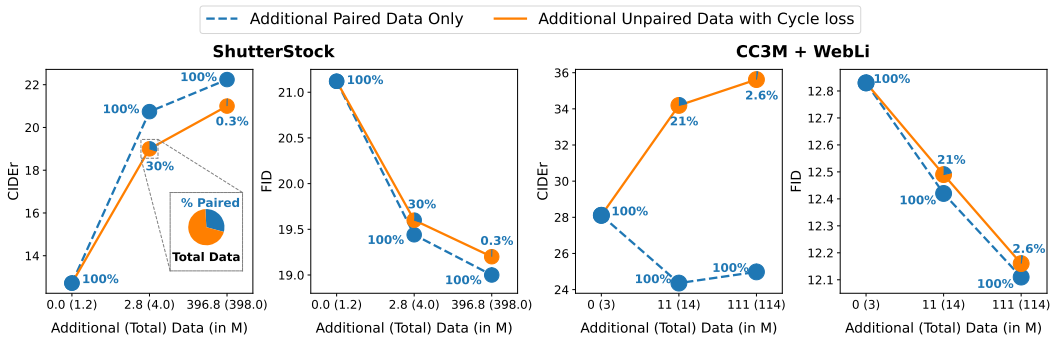

Figure 5: How ITIT's performance scales with the total amount of data used (x-axis). The baseline (T2I + I2T) in blue is trained entirely with increasing amounts of paired data. ITIT (orange) is trained with an increasing amount of unpaired data using cycle loss, while keeping the total amount of data equal for both curves. For example, the rightmost point with Shutterstock uses 1.2M image-text pairs and 396.8M unpaired samples (half as unpaired image and half as unpaired text) for ITIT with cycle loss, and 398M image-text pairs for the baseline. *Left*: Shutterstock data as both paired and unpaired. *Right*: CC3M as paired data, and varying fractions of WebLI as additional paired / unpaired data.

Next, we evaluate how ITIT's performance scales w.r.t. the total amount of data used. We first train a model with 1.2M paired image-text Shutterstock data. We then evaluate the effect of training models on adding increasing amounts of additional paired data vs. adding increasing amounts of unpaired data with cycle loss, keeping the total amount of data the same for both. As expected, we see in Figure 5 that performance scales up with additional paired data. Surprisingly, however, additional unpaired data exhibits similar scalability as paired. In fact, we can achieve 19.2 FID and 21.0 CIDEr with only 1.2M paired and 396.8M unpaired examples, which is very competitive with 19.0 FID and 22.2 CIDEr using 398M paired examples only. This experiment thus demonstrates that when scaling up training data, practitioners can rely on only adding unpaired examples using our method and achieve similar performance as paired data without the extra manual effort required to collect it.

We repeat the above experiment in a more realistic setting, where the small-scale paired dataset can contain high-quality image-text pairs but a large-scale paired dataset has much lower quality. For this, we use the high-quality CC3M as the paired dataset, and the much larger WebLI as the low-quality unpaired dataset. As before, we start with a model trained on 3M paired examples (from CC3M), and add additional training data from WebLI in paired (blue) or unpaired (orange) form. As shown in Figure 5, right pair, adding low-quality image-text pairs harms image captioning performance severely for the fully-paired case. However, the ITIT regime is not affected by this low quality and scales similarly as before. This demonstrates that our method is robust to low data quality in large datasets, and can in fact be used to achieve significantly better performance in settings when paired data is present but of low quality.

4.3 Comparison to Prior Work

In Table 1, we compare ITIT with state-of-the-art image-to-text and text-to-image models on the commonly used MS-COCO benchmark. As shown, all SOTA methods rely heavily on training on a large corpus of paired image-text data. ITIT, however, is trained with only 3M paired examples (CC3M), and an additional 55M unpaired image and text examples each (WebLI). Despite this, it beats many other methods trained on much more data for text-to-image generation (FID). For I2T, it beats methods using a comparable amount of data (highlighted in green), and achieves performance competitive with other SOTA methods. We find that the pre-training data (both the mixture and the size) also makes a difference to CIDEr score. For example, GIT (Wang et al., 2022a) achieves only 89.0 CIDEr fine-tuning performance on COCO captions when trained from scratch with 10M image-text pairs, which is far from its reported performance (144.8) when trained with 800M image-text pairs. Our approach is orthogonal to dataset mixture considerations, and we believe that scaling data size and variety will further enhance FID and CIDEr scores. We leave this to future work.

Table 1: Quantitative comparison with state-of-the-art text-to-image and image-to-text models on MS-COCO. The image-captioning performance is evaluated on the COCO Karpathy split, and the text-to-image generation FID is evaluated on 30K COCO images. † denotes our re-implementation. We highlight in green other models that use comparable amounts of paired data. Note that the GIT (CLIP) model uses a CLIP (Radford et al., 2021) encoder pre-trained with 400M image-text pairs.

Methods	#params	#paired data	#unpaired data	FID↓	CIDEr↑ (zs)	CIDEr↑ (ft)
T2I						
StableDiffusion (Rombach et al., 2022)	800M	400M	-	12.60	-	-
GLIDE (Nichol et al., 2021)	5B	250M	-	12.24	-	-
Make-A-Scene (Gafni et al., 2022)	4B	35M	-	11.84	-	-
DALL-E 2 (Ramesh et al., 2022)	3.5B	650M	-	10.39	-	-
PARTI (Yu et al., 2022b)	750M	5000M	-	10.71	-	-
Muse-512 (Chang et al., 2023)	3B	860M	-	7.88	-	-
Muse† (Chang et al., 2023)	750M	3M	-	23.7	-	-
I2T						
BLIP (Li et al., 2022)	446M	129M	-	-	-	136.7
$\text{SimVLM}_{\text{base}}$ (Wang et al., 2022b)	-	1100M	365M T	-	24.0	134.8
$\text{SimVLM}_{\text{huge}}$ (Wang et al., 2022b)	∼1.4B	1100M	365M T	-	32.2	143.3
GIT (CLIP) (Wang et al., 2022a)	681M	800M	-	-	-	144.8
GIT_B (scratch) (Wang et al., 2022a)	129M	10M	-	-	-	89.0
T2I+I2T						
CoBIT-Base (You et al., 2023)	626M	5200M	-	10.35	43.0	135.4
CoBIT-Large (You et al., 2023)	1091M	5200M	-	9.37	44.8	139.5
CM3Leon (Yu et al., 2023)	7B	340M	-	4.88	61.6	-
ITIT-B	221M	3M	55M I+55M T	13.4	32.1	103.5
ITIT-L	487M	3M	55M I+55M T	12.0	35.1	116.4
ITIT-H	868M	3M	55M I+55M T	10.4	38.2	125.3

4.4 ABLATIONS

In Table 2, we ablate the effectiveness of the four components of ITIT: T2I, I2T, T2I2T, and I2T2I. As shown in rows 1-3, combining T2I and I2T training in our framework already improves image captioning performance. This is likely because the T2I training alleviates the overfitting problem of I2T training, as shown in GIT (Wang et al., 2022a).

As before (Figure 5), we can see in row 4 that combining CC3M and WebLI improves text-to-image generation, but harms image captioning performance. This is because of the lower image-text pairing quality of WebLI compared to CC3M. The remaining rows demonstrate that the cycle loss alleviates this by using WebLI as unpaired data and does not depend on its image-text pairing quality. It is thus more generalizable to large-scale image-text datasets.

Next, rows 5-7 are naive baselines for using unpaired image or text data during generative vision-language training. We can simply perform text-to-text (T2T) autoregressive training without conditioning on images, which has been explored in some prior works (Wang et al. (2022b)). Similarly, we can perform image-to-image (I2I) reconstructive training without conditioning on text. Such baselines do improve the performance over not using any paired data (row 3).

We consider an ablation where the gradient of the cycle consistency loss is backpropagated up until the argmax step. Hence, only half of the cycle is trained. In fact, this is equivalent to first synthesizing an image counterpart from unpaired text and then using it as a pseudo image-text pair to train the I2T model (similarly for T2I). Rows 8-10 show that the half-cycle loss achieves much better performance than non-cycle baselines.

Finally, rows 11-14 show the performance of the full cycle ITIT training. Although T2I2T favors image captioning while I2T2I favors text-to-image generation, they both show significant improvement in text-to-image generation and image captioning. Moreover, row 14 demonstrates that such two cycle losses can be combined to further improve performance. Additionally, we can see that the full cycle loss beats the half-cycle baselines (row 8-10), demonstrating the effectiveness of the gradient estimation step.

Lastly, we find by comparing row 3 and 13 that the cycle consistency loss can slightly improve the performance even without any additional data. We believe this is because it forces better image-text alignment. However, comparing row 13 and 14 shows that the huge improvements in both text-to-image and image-to-text generation mainly stem from the usage of additional unpaired data.

Table 2: Quantitative comparison between different variants of ITIT on MS-COCO. All experiments use ITIT_{B} trained with 500K steps. We take 50% of WebLI data and use the images as our unpaired image data, and the other 50% of WebLI data and use the texts as our unpaired text data.

	T2I	I2T	T2I2T	I2T2I	paired data	unpaired text	unpaired image	FID↓	CIDEr↑
Paired data only									
1	✓	✗	✗	✗	CC3M	✗	✗	15.5	N/A
2	✗	✓	✗	✗	CC3M	✗	✗	N/A	19.0
3	✓	✓	✗	✗	CC3M	✗	✗	15.7	23.5
4	✓	✓	✗	✗	CC3M+WebLI	✗	✗	14.2	20.7
Paired+unpaired data, no cycle									
5	✓	✓	T2T	✗	CC3M	50% WebLI	✗	15.1	26.0
6	✓	✓	✗	I2I	CC3M	✗	50% WebLI	15.9	24.2
7	✓	✓	T2T	I2I	CC3M	50% WebLI	50% WebLI	15.6	28.5
Paired+unpaired data, half cycle									
8	✓	✓	Half	✗	CC3M	50% WebLI	✗	14.8	27.6
9	✓	✓	✗	Half	CC3M	✗	50% WebLI	14.7	24.8
10	✓	✓	Half	Half	CC3M	50% WebLI	50% WebLI	14.5	30.5
Paired+unpaired data, full cycle									
11	✓	✓	Full	✗	CC3M	50% WebLI	✗	14.6	28.4
12	✓	✓	✗	Full	CC3M	✗	50% WebLI	14.6	26.3
13	✓	✓	Full	Full	CC3M	CC3M	CC3M	15.4	24.4
14	✓	✓	Full	Full	CC3M	50% WebLI	50% WebLI	**14.3**	**31.1**

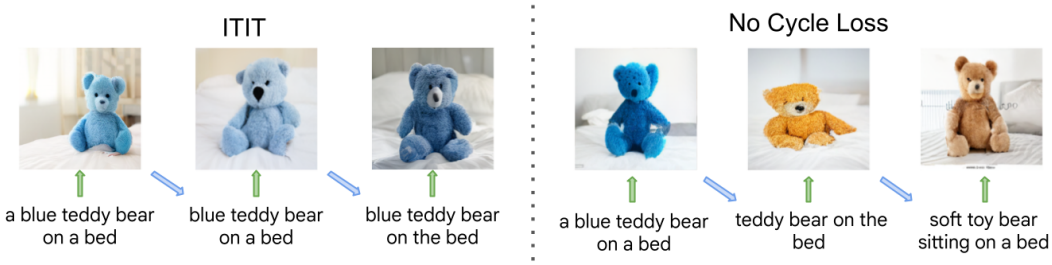

Figure 6: Iteratively generating text to image to text and so on. With ITIT, the generated results are more consistent than the results from a model trained without the cycle consistency loss.

4.5 Cycle-Generation Results

With a framework that can perform both image-to-text and text-to-image, we can easily perform cycle-generation, as shown in Figure 6. With ITIT training, the cycle generation often keeps the same semantics as the input text prompt. On the other hand, without the cycle consistency training, the cycle generation misses the "blue" semantics after the first cycle. This demonstrates that our cycle consistency training not only enables integrating unpaired image and text data into generative vision-language training, but also improves image-text alignment for both image-to-text and text-to-image generation. We include a number of results of image and text generation in Appendix A (Figures 1 through 4).

5 Discussion

We propose ITIT, a novel training scheme that for the first time incorporates unpaired images and text into generative vision-language training. Through extensive ablations, we demonstrate the effectiveness of both the T2I2T cycle and I2T2I cycle in improving text-to-image and image-to-text generation performance. As a result, ITIT achieves performance competitive with state-of-the-art vision-language generative models, but with only 3 million paired image-text samples. Our method can be used even when paired image-text data is present, and is especially helpful when the pairing quality is low. Future directions include scaling ITIT to larger unpaired image and text data and model sizes, and utilizing more diverse datasets.

REFERENCES

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

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.

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. In *Int. Conf. on Learning Representations (ICLR)*, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4125–4134, 2019.

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 *European Conference on Computer Vision*, pp. 89–106. Springer, 2022.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016. URL `http://arxiv.org/abs/1609.03677`.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Satya Krishna Gorti and Jeremy Ma. Text-to-image-to-text translation using cycle consistent adversarial networks. *arXiv preprint arXiv:1808.04538*, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

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*, pp. 3128–3137, 2015.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

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*, 2023a.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023b.

Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, apr 2022. doi: 10.1109/lra.2022.3145053. URL https://doi.org/10.1109%2Flra.2022.3145053.

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.

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*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

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*, pp. 8821–8831. PMLR, 2021.

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.

Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. Technical report, Google, 2019.

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*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

SentencePiece. Unsupervised text tokenizer for neural network-based text generation. https://github.com/google/sentencepiece, 2023.

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

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Shutterstock. Shutterstock. https://www.shutterstock.com/, 2023.

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

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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*, pp. 4566–4575, 2015.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022a.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2022b.

Haoxuan You, Mandy Guo, Zhecan Wang, Kai-Wei Chang, Jason Baldridge, and Jiahui Yu. Cobit: A contrastive bi-directional image-text generation model, 2023.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

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*, 2022a.

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*, 2(3):5, 2022b.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023.

Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1426–1433, 2010. doi: 10.1109/CVPR.2010.5539801.

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

Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. *CoRR*, abs/1604.05383, 2016. URL http://arxiv.org/abs/1604.05383.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.