Variational Distribution Learning for Unsupervised Text-to-Image Generation

Minsoo Kang[1*]   Doyup Lee[3]   Jiseob Kim[3]   Saehoon Kim[3]   Bohyung Han[1,2]

[1]ECE & [2]IPAI, Seoul National University   [3]Kakao Brain

{kminsoo, bhhan}@snu.ac.kr   {doyup.lee, jiseob.kim, shkim}@kakaobrain.com

Abstract

We propose a text-to-image generation algorithm based on deep neural networks when text captions for images are unavailable during training. In this work, instead of simply generating pseudo-ground-truth sentences of training images using existing image captioning methods, we employ a pretrained CLIP model, which is capable of properly aligning embeddings of images and corresponding texts in a joint space and, consequently, works well on zero-shot recognition tasks. We optimize a text-to-image generation model by maximizing the data log-likelihood conditioned on pairs of image-text CLIP embeddings. To better align data in the two domains, we employ a principled way based on a variational inference, which efficiently estimates an approximate posterior of the hidden text embedding given an image and its CLIP feature. Experimental results validate that the proposed framework outperforms existing approaches by large margins under unsupervised and semi-supervised text-to-image generation settings.

1. Introduction

Recent advances in text-to-image (T2I) generation techniques [6, 8, 17, 18, 21, 26–29, 36] have shown promising results by employing generative adversarial networks [9], autoregressive models [33], or diffusion models [11, 32] to synthesize images based on their text captions. However, these approaches require a paired dataset that consists of images and their corresponding text captions, and, consequently, incur significant annotation costs, especially for labeling image captions. To alleviate this limitation, unsupervised learning methods for T2I generation have recently drawn attention to the computer vision community, where the models learn to generate images without paired text captions.

Existing T2I models [1, 2, 34, 38] based on unsupervised learning exploit Contrastive Language-Image Pretraining (CLIP) [25] to sidestep the absence of text captions dur-

ing training. Specifically, after a text embedding is estimated using a given image embedding, the T2I model is trained to synthesize an image conditioned on the estimated text embedding. However, although image and text embeddings extracted by CLIP are not accurately aligned, existing approaches assume that the distinction is ignorable [34] or simple to recover by just adding Gaussian noises [38] without considering the underlying structure of text embeddings. Thus, those algorithms may suffer from large discrepancies between true and estimated text embeddings at both training and testing.

To tackle the challenge, we propose a variational distribution learning technique for unsupervised T2I generation, where the lower-bound of the data log-likelihood is maximized in a principled way. Specifically, we first regard a text embedding as a hidden random variable while an image and its CLIP embedding are observable random variables. Then, we decompose the variational lower-bound into three parts: 1) the similarity between the text embedding prior and posterior, 2) the log-likelihood of the image embedding given the text embedding, 3) the log-likelihood of the image given the image and text embeddings in the trained T2I model. Since the lower-bound formulation enforces the matching between the prior and posterior distributions of text embedding, our method achieves a more accurate estimation of the embedding and reduces the discrepancy between the true and estimated embeddings.

For the optimization of the variational objective, we employ a two-stage training strategy for T2I models. In the first stage, we learn an encode-decoder architecture that takes the image embedding as an input and the estimated text embedding as a latent bottleneck. Then, our network estimates two conditional distributions of CLIP embeddings, one for the variational distribution of the text embedding given the image embedding and the other for the model distribution of the image embedding given the text embedding. The parameters of the two distributions are obtained from the first two terms in the variational lower-bound objective. Note that we relax the Kullback-Leibler (KL) divergence term in the training objective of the first stage to an adversarial training loss, specifically, the Jensen-Shannon divergence. Since

---

*This work was partly done during an internship at Kakao Brain.

the KL divergence is only tractable for a confined family of distributions, this relaxation allows more flexible choices for the conditional and the prior distributions. In the second stage, a T2I model learns the conditional distribution of the images given the estimated text embeddings and the image features. Altogether, the proposed method achieves outstanding performance on widely adopted datasets [19, 31]. The main contributions of our work are summarized below:

- We propose a principled approach for unsupervised and semi-supervised text-to-image generation tasks based on a variational inference technique.

- We theoretically show that our method considers the underlying structure of text embeddings, which can eventually lead to better generalization performance.

- We empirically confirm that the proposed algorithm outperforms the existing methods by large margins.

The rest of our paper is organized as follows. Section 2 overviews the related work about unsupervised training methods in text-to-image generation. Section 3 describes the main idea of our approach while Sections 4 and 5 discuss the procedures of the first and second training stages, respectively. The experimental results are presented in Section 6, and we finally conclude this paper in Section 7.

2. Related Work

Text-to-image generative models have shown astonishing performance via learning with large-scale datasets composed of image-text pairs. Existing algorithms often represent each image with a sequence of discrete tokens and learn autoregressive [6, 8, 17, 27, 35] or bidirectional [18] transformers to generate high-resolution images given text inputs. Recently, the introduction of diffusion models [5, 11–13] has paved the way to learn large-scale T2I models and generate high-quality images conditioned on text.

For training T2I models without captions describing images, previous approaches [1, 34, 38] typically exploit the pretrained CLIP [25] to approximate missing text captions. Specifically, CLIP-GEN [34] assumes that a CLIP image embedding is perfectly aligned with the corresponding text embedding, and utilizes the image embedding as a proxy of its text embedding for T2I generation. On the other hand, LAFITE [38] adds a Gaussian random noise to the image embedding for estimating the unknown true text embedding. However, these algorithms fail to consider the underlying structure of the text embeddings, which eventually results in the imprecise approximation of text embedding. Retrieval-based approaches [1, 2] employ image features similar to CLIP-GEN [34] for text conditions while k-nearest image embeddings are additionally utilized for the construction of the conditions; note that these approaches
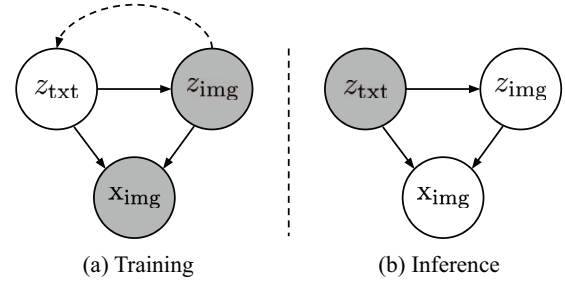

(a) Training   (b) Inference

Figure 1. Probabilistic graphical models of the proposed method, where shaded nodes represent observable random variables while unshaded nodes represent hidden random variables.

are orthogonal to our method and can be combined with the proposed method to further enhance generation performance.

On the other hand, we propose a principled framework relying on a variational inference to effectively reduce the discrepancy between the true embedding employed during inference and the approximated one drawn by the variational distribution used for training.

3. Main Framework

This section presents the existing unsupervised training approaches for T2I models. Then, we formulate our variational inference framework for unsupervised training that effectively reduces the discrepancy between true and estimated text embeddings. Figure 1 illustrates the graphical model of the data generating process in our approach.

3.1. Unsupervised Training of T2I Models

Unsupervised training of T2I models learns to generate images without textual annotations while the model generates high-quality images based on given captions at inference time. Conventional supervised training of T2I models exploits both an image, $\mathbf{x}_{\mathrm{img}}$, and its textual caption, $\mathbf{x}_{\mathrm{txt}}$, to estimate a conditional distribution $p(\mathbf{x}_{\mathrm{img}}|\mathbf{x}_{\mathrm{txt}})$. However, unsupervised training assumes that the text condition $\mathbf{x}_{\mathrm{txt}}$ is unavailable during training. Existing approaches estimate a latent text representation $\mathbf{z}_{\mathrm{txt}}$ and formulate the T2I model as a task to derive the conditional distribution $p(\mathbf{x}_{\mathrm{img}}|\mathbf{z}_{\mathrm{txt}})$. In other words, an encoder $E(\cdot)$ approximates the text condition, i.e., $\hat{\mathbf{z}}_{\mathrm{txt}} = E(\mathbf{x}_{\mathrm{img}})$, during training and $\hat{\mathbf{z}}_{\mathrm{txt}}$ is replaced with a representation of a given text sentence $\mathbf{x}_{\mathrm{txt}}$ at inference time.

The previous studies commonly incorporate the pretrained CLIP model [25], which consists of two separated encoders for images $f_{\mathrm{img}}(\cdot)$ and texts $f_{\mathrm{txt}}(\cdot)$. The vision-language model is learned to make a pair of image and text embeddings, denoted respectively by $f_{\mathrm{img}}(\mathbf{x}_{\mathrm{img}})$ and $f_{\mathrm{txt}}(\mathbf{x}_{\mathrm{txt}})$, have a high cosine similarity. CLIP-GEN [34] and

retrieval-based models [1, 2] approximate the text condition using the CLIP image embedding, $\hat{\mathbf{z}}_{\text{txt}} \approx f_{\text{img}}(\mathbf{x}_{\text{img}})$, during training. Since the two embeddings are not exactly aligned, LAFITE [38] additionally adds a perturbation to the CLIP image embedding and approximates the text condition $\mathbf{z}_{\text{txt}}$ as

$$\hat{\mathbf{z}}_{\text{txt}} \approx \text{Normalize}(f_{\text{img}}(\mathbf{x}_{\text{img}}) + \xi \|f_{\text{img}}(\mathbf{x}_{\text{img}})\| \boldsymbol{\epsilon}/\|\boldsymbol{\epsilon}\|), \quad (1)$$

where $\|\cdot\|$ and $\text{Normalize}(\cdot)$ are the Euclidean norm and an operator to divide the input by its magnitude while ξ is a hyperparameter and $\boldsymbol{\epsilon}$ is a random noise drawn from the Gaussian distribution, $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Although the previous methods are proposed to approximate the absent text condition, we observe that they are insufficient to reduce the gap between the training and the inference environments of T2I models.

3.2. Variational Inference for Training T2I Models

We aim to formulate a variational inference framework for unsupervised training of T2I models to reduce the discrepancy between training and inference while improving the performance of T2I models. Given CLIP image and text embeddings, $\mathbf{z}_{\text{img}} = f_{\text{img}}(\mathbf{x}_{\text{img}})$ and $\mathbf{z}_{\text{txt}} = f_{\text{txt}}(\mathbf{x}_{\text{txt}})$, our T2I generation model is defined as

$$p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}}), \quad (2)$$

where $\theta^{\mathbf{x}_{\text{img}}}$ is a set of parameters in the T2I model. Note that the text representation $\mathbf{z}_{\text{txt}}$ is unavailable during training while the image embedding $\mathbf{z}_{\text{img}}$ is absent for inference as illustrated in Figure 1. Thus, the challenges lie in the precise approximation of the text embedding $\hat{\mathbf{z}}_{\text{txt}}$ based on $\mathbf{z}_{\text{img}}$ during training while approximating $\hat{\mathbf{z}}_{\text{img}}$ given $\mathbf{z}_{\text{txt}}$ at inference.

Our approach naturally maximizes the marginal log-likelihood $\log p_{\theta}(\mathbf{x}_{\text{img}})$ with respect to θ, a set of parameters of our data generating process, without the observation of the text embedding $\mathbf{z}_{\text{txt}}$ during training. The log-likelihood $\log p_{\theta}(\mathbf{x}_{\text{img}})$ is computed by marginalizing out $\mathbf{z}_{\text{txt}}$ as

$$\log p_{\theta}(\mathbf{x}_{\text{img}}) = \log p_{\theta}(\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}})$$
$$= \log \int p_{\theta}(\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}}) d\mathbf{z}_{\text{txt}}. \quad (3)$$

Unfortunately, it is intractable to compute the marginal log-likelihood either directly or through the estimation of the posterior, $p_{\theta}(\mathbf{z}_{\text{txt}}|\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}})$. Hence, we employ a variational inference technique that approximates the true posterior $p_{\theta}(\mathbf{z}_{\text{txt}}|\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}})$ using a variational distribution $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$ parametrized by $\phi^{\mathbf{z}_{\text{txt}}}$ under the assumption of $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}}) = q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$.

To learn the true posterior distribution via the variational posterior, we equivalently maximize the lower bound on the log-likelihood, which is given by

$$\log p_{\theta}(\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}})$$
$$\geq \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})} \left[\log \frac{p_{\theta}(\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})}{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})} \right], \quad (4)$$

where the lower bound in (4) is defined as $\mathcal{L}_{\text{ELBO}}$. We achieve the tight lower bound when the variational distribution is exactly same as the true posterior.

ELBO By factorizing $p_{\theta}(\mathbf{x}_{\text{img}}, \mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})$ into the product of $p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})$, $p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})$, and $p(\mathbf{z}_{\text{txt}})$ according to the probabilistic graphical model in Figure 1, $\mathcal{L}_{\text{ELBO}}$ is decomposed as follows:

$$\mathcal{L}_{\text{ELBO}} = -D_{\text{KL}}(q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})||p(\mathbf{z}_{\text{txt}}))$$
$$+ \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})}[\log p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})]$$
$$+ \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})}[\log p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})]. \quad (5)$$

We reformulate our task as the maximization of $\mathcal{L}_{\text{ELBO}}$, which involves the optimization of the T2I model in (2) under the unsupervised setting. Specifically, the maximization of the first term in (5) encourages the estimated sample $\hat{\mathbf{z}}_{\text{txt}}$ drawn from $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$ to lie on the structure of $\mathbf{z}_{\text{txt}}$ by minimizing the KL divergence. Also, the second term learns $p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})$ to reconstruct $\mathbf{z}_{\text{img}}$ based on the given $\mathbf{z}_{\text{txt}}$ from $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$, where $p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})$ is employed to estimate $\mathbf{z}_{\text{img}}$ in (2) at inference. Finally, the third term implies training T2I models without text captions, where $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$ estimates the absent text embedding $\mathbf{z}_{\text{txt}}$ based on $\mathbf{z}_{\text{img}}$. The details about how to optimize the model parameters are discussed in the next section.

To maximize $\mathcal{L}_{\text{ELBO}}$ of (5) in practice, we employ the following two-stage optimization procedure:

1. **Fix $\{\theta^{\mathbf{x}_{\text{img}}}\}$ and optimize $\theta^{\mathbf{z}_{\text{img}}}$ and $\phi^{\mathbf{z}_{\text{txt}}}$.**

   With the parameter $\{\theta^{\mathbf{x}_{\text{img}}}\}$ fixed, maximize the objective in (5) with respect to $\theta^{\mathbf{z}_{\text{img}}}$ and $\phi^{\mathbf{z}_{\text{txt}}}$.

2. **Fix $\{\theta^{\mathbf{z}_{\text{img}}}, \phi^{\mathbf{z}_{\text{txt}}}\}$ and optimize $\theta^{\mathbf{x}_{\text{img}}}$.**

   With the parameters $\{\theta^{\mathbf{z}_{\text{img}}}, \phi^{\mathbf{z}_{\text{txt}}}\}$ fixed, maximize the objective in (5) with respect to $\theta^{\mathbf{x}_{\text{img}}}$.

In other words, we first train the generative and variational parameters to approximate $\mathbf{z}_{\text{img}}$ and $\mathbf{z}_{\text{txt}}$, and then train a T2I model under the unsupervised setting. We present the detailed description of the first and second stages in Section 4 and 5, respectively.

4. Variational Distribution Learning

This section describes the first stage of our algorithm based on variational distribution learning (VDL). VDL approximates the posterior of text features given image embeddings, $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$, via an adversarial training and

reconstructs the image embeddings given the text features using $p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})$.

4.1. Sampling

We first discuss the sampling procedure in VDL that involves the sampling of $\hat{\mathbf{z}}_{\text{txt}}$ for training to estimate the text embeddings and the sampling of $\hat{\mathbf{z}}_{\text{img}}$ for inference to obtain the image embeddings. Also, we describe the sampling process of the text prior $\mathbf{z}'_{\text{txt}}$, which is only required for training.

4.1.1 Generating Samples for Text Embedding

Let $G(\cdot)$ with parameters of $\phi^{\mathbf{z}_{\text{txt}}}$ be an encoder implemented with a multilayer perceptron, Using the encoder, we draw a sample $\hat{\mathbf{z}}_{\text{txt}}$ from an implicitly defined variational distribution $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$ as follows:

$$
\begin{aligned}
\hat{\mathbf{z}}_{\text{txt}} &\sim S_{\text{VDL}}(z_{\text{img}}, G, r) \\
&:= \text{Normalize}\left(\mathbf{z}_{\text{img}} + r \cdot \frac{G(\mathbf{z}_{\text{img}})}{\|G(\mathbf{z}_{\text{img}})\|}\right),
\end{aligned} \tag{6}
$$

where $S_{\text{VDL}}(\cdot, \cdot, \cdot)$ denotes our sampling strategy and $r \geq 0$ is a hyperparameter. The following proposition validates the high cosine similarity between the estimated text sample $\hat{\mathbf{z}}_{\text{txt}}$ and the image feature $\mathbf{z}_{\text{img}}$ that has a similar representation with the unknown text embedding $\mathbf{z}_{\text{txt}}$.

Proposition 1. *Let $\hat{\mathbf{z}}_{txt}$ be a sample obtained by the proposed sampling strategy S_{VDL} defined in (6) based on $\mathbf{z}_{img}$. Then, the following inequality always holds for $G(\cdot; \phi^{\mathbf{z}_{txt}})$ with arbitrary values of its parameter $\phi^{\mathbf{z}_{txt}}$:*

$$
\hat{\mathbf{z}}_{txt}^T \mathbf{z}_{img} \geq \sqrt{1 - r^2},
$$

where $r < 1$.

We provide the proof of this proposition in the supplementary document. Proposition 1 implies that a sample $\hat{\mathbf{z}}_{\text{txt}}$ drawn by (6) is guaranteed to be close to $\mathbf{z}_{\text{img}}$ by setting r to a sufficiently small number in $[0, 1]$. The proposed sampling strategy also results in the acceleration of the optimization procedure by constraining the search space of the variational distribution. Thanks to the optimization procedure, $\hat{\mathbf{z}}_{\text{txt}}$ is expected to achieve high cosine similarity with $\mathbf{z}_{\text{txt}}$, which is also empirically confirmed in Section 6.

4.1.2 Generating Samples for Image Embedding

During training, we reconstruct the image embedding samples to compute the second expectation in (5) while we obtain the embeddings to synthesize the images during inference. Similar to the sampling process of $q(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$, we reconstruct the image embedding based on the text feature

$\mathbf{z}_{\text{txt}}$ via the sampling process of $p(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})$ using a decoder $F(\cdot)$ with $\theta^{\mathbf{z}_{\text{img}}}$, which is given by

$$
\begin{aligned}
\hat{\mathbf{z}}_{\text{img}} &\sim S_{\text{VDL}}(\mathbf{z}_{\text{txt}}, F, r) \\
&= \text{Normalize}\left(\mathbf{z}_{\text{txt}} + r \cdot \frac{F(\mathbf{z}_{\text{txt}})}{\|F(\mathbf{z}_{\text{txt}})\|}\right),
\end{aligned} \tag{7}
$$

where $\hat{\mathbf{z}}_{\text{img}}$ is a reconstructed image embedding. According to Proposition 1, $\hat{\mathbf{z}}_{\text{img}}$ is also expected to have a high cosine similarity with $\mathbf{z}_{\text{txt}}$.

4.1.3 Generating Prior Samples for Text Embedding

For the first stage of training, we require samples from the prior to minimize the difference between the variational and text prior distributions with respect to the parameters in the encoder $G(\cdot)$. Refer to Section 4.2 for the detailed optimization procedure. We draw a sample $\mathbf{z}'_{\text{txt}}$ from the prior distribution using a text corpus, which is available online. Note that we do not employ an additional dataset consisting of image-text pairs, but randomly choose a sentence $\mathbf{x}'_{\text{txt}}$ from the text corpus. Using the sampled sentence, the representation of the prior sample is obtained by

$$
\mathbf{z}'_{\text{txt}} = f_{\text{txt}}(\mathbf{x}'_{\text{txt}}). \tag{8}
$$

In our framework, the pretrained CLIP encoder is fixed to reduce the computational burden.

4.2. Training

4.2.1 Robust Objective

The optimization of (5) with respect to $\theta^{\mathbf{z}_{\text{img}}}$ and $\phi^{\mathbf{z}_{\text{txt}}}$ is equivalent to solve the following minimization problem because $\log p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})$ is irrelevant. Hence, the objective function for the first stage is given by

$$
\begin{aligned}
\min_{\theta^{\mathbf{z}_{\text{img}}}, \phi^{\mathbf{z}_{\text{txt}}}} & D_{\text{KL}}(q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})\|p(\mathbf{z}_{\text{txt}})) \\
& - \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})}[\log p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})].
\end{aligned} \tag{9}
$$

However, the KL divergence is generally intractable unless the two distributions belong to specific families of probability distributions, *e.g.*, Gaussian distribution. Such restricted distributions are different from the true distributions of the CLIP features, which only have non-zero densities at the surface of the unit hypersphere in the feature space. Therefore, the use of the restricted distributions for modeling the variational and prior distributions would cause high approximation errors [4]. To reduce the distribution gap, the von Mises-Fisher distribution can be employed but this probability density involves the Bessel function, which also leads to the intractable KL divergence.

To address the issue, we alternatively minimize the difference between the two distributions using the Jensen-Shannon (JS) divergence that also enforces the two distributions to become identical via its minimization. Contrary

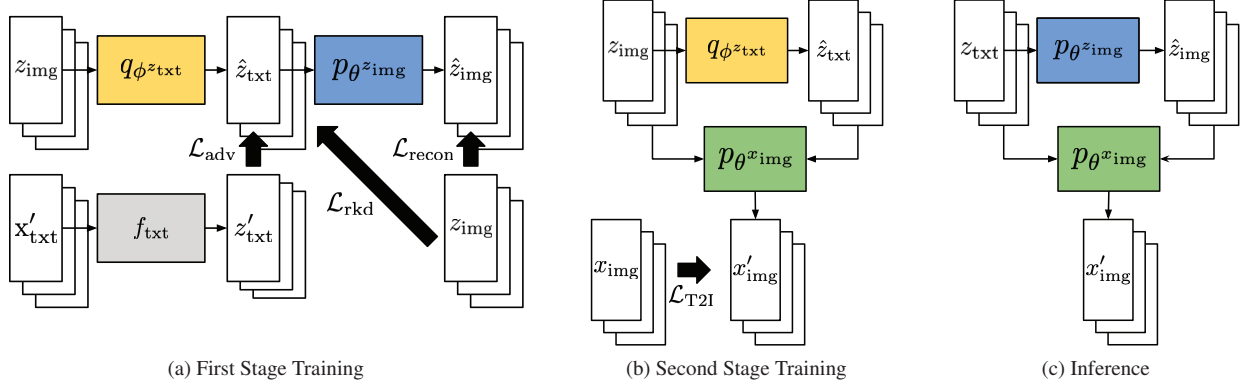(a) First Stage Training (b) Second Stage Training (c) Inference

Figure 2. Overview of the proposed method. We illustrate the first and second stage training procedures and then depict the inference step.

to the KL divergence, the JS divergence is always bounded and free from density assumption, which allows us to use implicit generative networks for flexible modeling of the two distributions. We observe that replacing the KL divergence with the JS divergence is more effective, which will be empirically validated in Section 6.6.1. By using the property discussed in [9, 22], the objective is reformulated as the following minimax game, which is given by

$$
\min_{\theta^{\mathbf{z}_{\text{img}}}, \phi^{\mathbf{z}_{\text{txt}}}} \max_{\rho^D} \mathbb{E}_{p(\mathbf{z}_{\text{txt}})}[\log D(\mathbf{z}_{\text{txt}})]
$$
$$
+ \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})} \left[\log(1 - D(\mathbf{z}_{\text{txt}}))\right]
$$
$$
- \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})} \left[\log p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})\right], \quad (10)
$$

where $D(\cdot)$ is a discriminator parametrized by ρ^D. The last term of (10) denoted by $\mathcal{L}_{\text{recon}}$ is expressed as

$$
\mathcal{L}_{\text{recon}} := \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})} \left[-\log p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})\right]
$$
$$
= \frac{1}{2\sigma^2} \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})} \left[\|\mathbf{z}_{\text{img}} - \hat{\mathbf{z}}_{\text{img}})\|^2\right], \quad (11)
$$

where we employ the ℓ_2 loss for reducing the negative log-likelihood with a balancing factor σ.

4.2.2 Relational Representation Transfer

In addition, we encourage the approximate text samples $\hat{\mathbf{z}}_{\text{txt}}$ to mimic the correlation of the observed image embedding samples $\mathbf{z}_{\text{img}}$ in order to mitigate the challenge posed by the lack of supervision. The intuition behind this strategy is that the structural relation of text embeddings will resemble that of image representations. For example, the two text embeddings should be located close if the image representations are similar, and vice versa. To impose the constraint on the text embedding samples, we employ a relational knowledge distillation framework [23], which makes a student mimic the relations among data embeddings given by a teacher. In

our framework, we view image embeddings as teacher samples while text embeddings are regarded as student ones. Therefore, we additionally minimize the relational distillation loss, which is given by

$$
\mathcal{L}_{\text{rkd}} := \mathbb{E}[\ell_\delta(\psi_A(\mathbf{z}_{\text{img}}^i, \mathbf{z}_{\text{img}}^j, \mathbf{z}_{\text{img}}^k) - \psi_A(\hat{\mathbf{z}}_{\text{txt}}^i, \hat{\mathbf{z}}_{\text{txt}}^j, \hat{\mathbf{z}}_{\text{txt}}^k))], \quad (12)
$$

where the expectation is taken over any triplet image embeddings $(\mathbf{z}_{\text{img}}^i, \mathbf{z}_{\text{img}}^j, \mathbf{z}_{\text{img}}^k)$ and the corresponding samples drawn by the variational distribution over $(\hat{\mathbf{z}}_{\text{txt}}^i, \hat{\mathbf{z}}_{\text{txt}}^j, \hat{\mathbf{z}}_{\text{txt}}^k)$. In the above equation, $\ell_\delta(\cdot)$ and $\psi_A(\cdot, \cdot, \cdot)$ are defined as

$$
\ell_\delta(a) := \begin{cases} \frac{1}{2}a^2, & \text{for } \|a\| \leq \delta, \\ \delta \cdot (\|a\| - \frac{1}{2}\delta), & \text{otherwise,} \end{cases} \quad (13)
$$
$$
\psi_A(\mathbf{z}^i, \mathbf{z}^j, \mathbf{z}^k) := \text{sim}(\mathbf{z}^i - \mathbf{z}^j, \mathbf{z}^i - \mathbf{z}^k), \quad (14)
$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between two vectors. We set the hyperparameter δ of the Huber loss $\ell_\delta(\cdot)$ to 1 instead of searching for it.

4.2.3 Total Objective

In summary, the final objective function of the first stage is given by

$$
\min_{\theta^{\mathbf{z}_{\text{img}}}, \phi^{\mathbf{z}_{\text{txt}}}} \max_{\phi^D} \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{recon}} + \lambda_{\text{rkd}} \mathcal{L}_{\text{rkd}}, \quad (15)
$$

where λ_{rkd} is a hyperparameter and $\mathcal{L}_{\text{adv}}$ denotes the first two terms of (10) constituting the adversarial loss. Figure 2a illustrates the optimization procedure of the first stage. In the case of the semi-supervised setting, we additionally employ the reconstruction loss for labeled examples, $\mathcal{L}_{\text{semi}}$, which enforces the variational text samples to mimic the true ones. The reconstruction loss is formally given by

$$
\mathcal{L}_{\text{semi}} := \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\hat{\mathbf{z}}_{\text{txt}}|\mathbf{z}_{\text{img}})} \left[\|\hat{\mathbf{z}}_{\text{txt}} - \mathbf{z}_{\text{txt}}\|_1\right], \quad (16)
$$

where $\|\cdot\|_1$ indicates the ℓ_1-norm.

23384

Table 1. Results of unsupervised text-to-image generation on the MS-COCO [19] and Conceptual Captions 3M [31] datasets using Style-GAN2 [14]. Captioning indicates a text-to-image generation baseline method relying on a state-of-the-art image captioning algorithm [37], where the results of the baselines are retrieved from [38]. Methods with asterisks * report the results of our reproduction. A bold-faced number denotes the best performance in each column while '–' indicates that the number is unavailable.

T2I Model	Dataset	Method	IS ($\uparrow$)	FID ($\downarrow$)	Sim_{txt} ($\uparrow$)	Sim_{img} ($\uparrow$)
StyleGAN2 [14]	MS-COCO [19]	Captioning [37]	15.83	56.36	-	-
		CLIP-GEN* [34]	16.94	58.63	0.3042	-
		LAFITE [38]	27.20	18.04	0.0965	-
		VDL (Ours)	**30.30**	**13.22**	**0.6104**	**0.7655**
	Conceptual Captions 3M [31]	CLIP-GEN* [34]	7.88	84.16	0.2896	-
		LAFITE* [38]	16.06	22.95	0.0912	-
		VDL (Ours)	**23.66**	**17.37**	**0.6237**	**0.7105**

5. Text-to-Image Generative Models

After training the first stage of VDL, we leverage the approximate posterior distribution of text features, $q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})$, for the unsupervised training of T2I models. For a T2I generation model with trainable parameters $\theta^{\mathbf{x}_{\text{img}}}$, we maximize (5) with respect to $\theta^{\mathbf{x}_{\text{img}}}$, which is equivalent to solve the following problem:

$$\max_{\theta^{\mathbf{x}_{\text{img}}}} \mathbb{E}_{q_{\phi^{\mathbf{z}_{\text{txt}}}}(\mathbf{z}_{\text{txt}}|\mathbf{z}_{\text{img}})}[\log p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})], \quad (17)$$

where $p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{txt}})$ is a T2I generative model. Figure 2b and 2c illustrate the inference procedures for training and testing, respectively.

Contrary to the existing methods such as CLIP-GEN [34] and LAFITE [38], which rely only on $\mathbf{z}_{\text{txt}}$ for conditional image generation, our approach utilizes $\mathbf{z}_{\text{img}}$ as well as $\mathbf{z}_{\text{txt}}$ for conditioning, resulting in better generalization performance thanks to the additional information for the image. In our framework, $\hat{\mathbf{z}}_{\text{img}}$ is reconstructed by the learned model, $p_{\theta^{\mathbf{z}_{\text{img}}}}(\mathbf{z}_{\text{img}}|\mathbf{z}_{\text{txt}})$, during inference, and then an image $\mathbf{x}_{\text{img}}$ is finally obtained from the optimized generator, $p_{\theta^{\mathbf{x}_{\text{img}}}}(\mathbf{x}_{\text{img}}|\hat{\mathbf{z}}_{\text{img}}, \mathbf{z}_{\text{txt}})$.

For a T2I model, we employ StyleGAN2 [14] while Latent Diffusion Model (LDM) [28] is also adopted, where the experimental results of LDM are provided in the supplementary material. When the StyleGAN2 synthesizes images conditioned on text, we replace each style vector $\mathbf{s}^u$ with its conditional counterpart $\mathbf{s}^c$ following [38], which is formally given by

$$\mathbf{s}^c = h([\mathbf{s}^u; g([\mathbf{z}_{\text{txt}}; \mathbf{z}_{\text{img}}])]), \quad (18)$$

where $h(\cdot)$ denotes an affine transform, $g(\cdot)$ is a neural network with two fully-connected layers, and $[\cdot; \cdot]$ is the concatenation operator.

6. Experiments

This section compares the proposed method referred to as VDL with existing approaches on the standard datasets under unsupervised and semi-supervised text-to-image generation settings, and analyzes the proposed components.

6.1. Datasets

We employ MS-COCO [19] and Conceptual Captions 3M [31] (CC3M) datasets, which are widely used for the evaluation of text-to-image generation tasks. As training and validation datasets, the MS-COCO dataset contains 82k and 40k images while the CC3M dataset consists of 3.3M and 16k examples, respectively. As a preprocessing, we resize all images to 256×256 pixels for training a T2I network while the images are resized to 224×224 pixels before feeding them into the CLIP image encoder. As text corpora, we use text captions in CC3M for MS-COCO and 3 million randomly sampled texts from Conceptual 12M [3] for CC3M. Note that we do not utilize the image-text pairs for the proposed algorithm like other unsupervised methods such as CLIP-GEN [34] and LAFITE [38].

6.2. Evaluation Metrics

We select the Fréchet Inception Distance (FID) [10] and Inception Score (IS) [30] to evaluate and compare the visual quality of generated images. We measure the two metrics following the experimental protocol of previous works [27, 38, 39] for fair comparisons. Additionally, we report Sim_{txt} and Sim_{img} using the validation dataset, which are calculated by the expected cosine similarity between the true and predicted text features and the similarity between the image and its inferred embeddings, respectively. Note that Sim_{img} cannot be measured for the baseline algorithms since they do not consider reconstructing the image feature. The discrepancy between the training and inference becomes lower when each of the similarities is higher.

6.3. Implementation Details

The proposed method is implemented with the official code of LAFITE[1] based on PyTorch [24]. For the encoder

---

[1]https://github.com/drboog/Lafite

Table 2. Results of semi-supervised text-to-image generation on the MS-COCO dataset using StyleGAN2. The 'Ratio' column shows the fractions of the labeled text captions in each dataset.

Model	Dataset	Method	Ratio	IS ($\uparrow$)	FID ($\downarrow$)	Sim_{txt} ($\uparrow$)	Sim_{img} ($\uparrow$)
StyleGAN2 [14]	MS-COCO [19]	LAFITE [38]	0.0	27.20	18.04	0.0965	–
		LAFITE* [38]	0.1	18.82	20.65	0.8340	–
		LAFITE* [38]	0.2	21.19	17.74	0.8373	–
		LAFITE* [38]	0.3	21.39	15.76	**0.8385**	–
		VDL (Ours)	0.0	30.30	13.22	0.6237	0.7105
		VDL (Ours)	0.1	32.81	**11.24**	0.7130	0.7536
		VDL (Ours)	0.2	**33.90**	**11.24**	0.7261	**0.7596**

$G(\cdot)$, decoder $F(\cdot)$, and discriminator $D(\cdot)$, we adopt networks consisting of 10 fully-connected layers with the leaky ReLU activations, where each hidden layer has 2048 units. For the discriminator, we add an R_1 regularization [20] for training stability, which suppresses the magnitudes of gradients. We use the Adam optimizer [15] for the three networks with an initial learning rate of 0.001 with a batch size of 512.

During the second training stage, we train a T2I model using StyleGAN2 [14] to follow the experimental protocol used in LAFITE [38] for fair comparisons. Specifically, we optimize StyleGAN2 using the Adam optimizer, and set a batch size to 64 and an initial learning rate to 2.5×10^{-3}. Also, R_1 regularization is also performed for the discriminator every 16 iterations to save training time.

6.4. Unsupervised Setting Results

Table 1 shows the text-to-image generation results on the MS-COCO [19] and CC3M [31] datasets under the unsupervised setting, where the image captions are not available during training. As demonstrated in the table, VDL achieves the best performance in terms of Sim_{txt}, FID, and IS by large margins both on the two datasets. Although the noise injection for text prediction in LAFITE [38] degrades Sim_{txt}, LAFITE achieves better T2I performance in terms of FID and IS than CLIP-GEN. This is partly because the regions corresponding to the noisy text features predicted by LAFITE are larger than the deterministic point features given by CLIP-GEN and LAFITE consequently has more chance to identify accurate text features within the region.

6.5. Semi-Supervised Setting Results

In Table 2, we present the performance of LAFITE [38] and VDL in semi-supervised learning scenarios on the MS-COCO dataset [19], where only a fraction of image and text pairs are accessible; the ratio of 0.0 indicates the unsupervised setting. According to the table, the proposed method outperforms LAFITE by large margins in terms of FID and IS. Also, VDL trained even in the unsupervised setting outperforms LAFITE with the ratio of 0.3, which implies that the proposed method is more annotation-efficient. As

Table 3. Comparison between the KL divergence with the JS divergence under the unsupervised setting on the MS-COCO dataset using StyleGAN2. KLD optimizes (9) with the assumption that the variational distribution follows the Gaussian distribution with Gaussian mixture models for the prior. DualKLD optimizes the dual representation of the KL objective in [7].

Method	IS ($\uparrow$)	FID ($\downarrow$)	Sim_{txt} ($\uparrow$)	Sim_{img} ($\uparrow$)
KLD	20.89	34.42	0.1396	0.2833
DualKLD	27.83	15.60	**0.6236**	0.7440
JSD (VDL)	**30.30**	**13.22**	0.6104	**0.7655**

more text captions become available, VDL obtains higher IS, Sim_{txt}, and Sim_{img} while FID is unfortunately saturated at an early stage. In terms of Sim_{txt} under semi-supervised settings, LAFITE outperforms VDL. That this is partly because LAFITE trains a learnable network on the paired data by making the predicted text embeddings close to the true ones. However, the strategy turns out to be ineffective for improving FID and even degrades the IS score.

6.6. Analysis

6.6.1 Jensen-Shannon Divergence

We study the effect of using the JS divergence instead of KL divergence under the unsupervised setting with the Style-GAN2 architecture on MS-COCO. As reported in Table 3, our strategy using the JS divergence is more effective than employing the KL divergence and its variation. Specifically, we compute the objective in (9) using two different ways. First, motivated by variational autoencoders [16], we model the variational and prior distributions with a Gaussian distribution and its mixture, respectively, and this approach is referred to as KLD. The other method denoted by DualKLD replaces the KL divergence with its dual form, the Donsker-Varadhan representation [7], which also performs the minimax optimization. In the case of KLD, the performance significantly degrades because the Gaussian assumptions for the variational and prior distributions are not effective; the Gaussian distribution has a non-zero density outside the unit-hypersphere, where CLIP text features

Figure 3. Qualitative results on the MS-COCO and Conceptual Captions 3M datasets using StyleGAN2. VDL generates visually higher-quality images than LAFITE and CLIP-GEN.

Table 4. Ablation study results on MS-COCO with StyleGAN2 under an unsupervised setting. VDL w/o S_{VDL} directly predicts a text feature and then normalizes it to locate at the unit-hypersphere without using S_{VDL} while VDL w/o $\mathcal{L}_{rkd}$ does not employ $\mathcal{L}_{rkd}$.

Method	IS ($\uparrow$)	FID ($\downarrow$)	Sim_{txt} ($\uparrow$)	Sim_{img} ($\uparrow$)
VDL w/o S_{VDL}	19.86	36.37	0.5498	0.5256
VDL w/o $\mathcal{L}_{rkd}$	28.45	15.75	**0.6128**	0.7637
VDL (Ours)	**30.30**	**13.22**	0.6104	**0.7655**

are not located. On the other hand, DualKLD, which is free from the distribution restriction, outperforms KLD although it is still worse than our approach. However, we observe that DualKLD is sensitive to the hyperparameter partly due to its unbounded property contrary to the JS divergence.

6.6.2 Component analysis

We analyze the contributions of the individual components in our approach. As presented in Table 4, S_{VDL} is helpful for improving Sim_{txt}, Sim_{img}, and T2I performance in unsupervised settings. Although $\mathcal{L}_{rkd}$ is conceptually irrelevant to reduce the discrepancy between the true and predicted text features, it improves generation performance by learning relational embeddings between two modalities.

6.6.3 Qualitative Results

Figure 3 visualizes generation results on the MS-COCO and CC3M datasets using CLIP-GEN [34], LAFITE [38], and

VDL. As illustrated in the figure, the proposed method successfully generates images based on given sentences with enhanced visual quality while the others sometimes fail to understand the overall meaning of text captions.

7. Conclusion

We presented an annotation-efficient method for text-to-image generation when image and text caption pairs are rarely available or text information is completely inaccessible. To address the challenge, we employ the off-the-shelf CLIP model to estimate hidden text features given observable images, where we rely on the CLIP's multi-modal joint embedding quality. To further improve the quality of text embedding, we approximate its intractable true posterior probability by exploiting the variational inference technique. Given the inferred features and their image embeddings, we learn a conditional generative model to reconstruct images. Experimental results verify that the proposed method achieves outstanding performance on the unsupervised and semi-supervised learning environments.

References

[1] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. KNN-Diffusion: Image Generation via Large-Scale Retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 1, 2, 3

[2] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022. 1, 2, 3

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing Web-Scale Image-Text Pre-training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021. 6

[4] Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality in Variational Autoencoders. In *ICML*, 2018. 4

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 2

[6] 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. In *NeurIPS*, 2021. 1, 2

[7] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic Evaluation of Certain Markov Process Expectations for Large Time. IV. *Communications on Pure and Applied Mathematics*, 1983. 7

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 2021. 1, 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NIPS*, 2014. 1, 5

[10] 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. In *NIPS*, 2017. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 1, 2

[12] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *JMLR*, 2022. 2

[13] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of Stylegan. In *CVPR*, 2020. 6, 7

[15] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 7

[16] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 7

[17] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive Image Generation using Residual Quantization. In *CVPR*, 2022. 1, 2

[18] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Draft-and-Revise: Effective Image Generation with Contextual RQ-Transformer. *arXiv preprint arXiv:2206.04452*, 2022. 1, 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2, 6, 7

[20] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs Do Actually Converge? In *ICML*, 2018. 7

[21] 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. In *ICML*, 2022. 1

[22] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *NIPS*, 2016. 5

[23] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *CVPR*, 2019. 5

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019. 6

[25] 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 *ICML*, 2021. 1, 2

[26] 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

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *ICML*, 2021. 1, 2, 6

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 6

[29] 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

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *NIPS*, 2016. 6

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018. 2, 6, 7

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based

generative modeling through stochastic differential equations. In *ICLR*, 2021. 1

[33] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *ICML*, 2016. 1

[34] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. *arXiv preprint arXiv:2203.00386*, 2022. 1, 2, 6, 8

[35] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual Synthesis Pre-Training for Neural Visual World Creation. *arXiv preprint arXiv:2111.12417*, 2021. 2

[36] 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. 1

[37] 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 *CVPR*, 2021. 6

[38] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards Language-Free Training for Text-to-Image Generation. In *CVPR*, 2022. 1, 2, 3, 6, 7, 8

[39] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dmgan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019. 6