Shifted Diffusion for Text-to-image Generation

Yufan Zhou [1]*, Bingchen Liu [2], Yizhe Zhu [2], Xiao Yang [2], Changyou Chen [1], Jinhui Xu [1]
[1] State University of New York at Buffalo [2] ByteDance
{yufanzho, changyou, jinhui}@buffalo.edu, {bingchenliu, yizhe.zhu, yangxiao.0}@bytedance.com

Abstract

We present Corgi, a novel method for text-to-image generation. Corgi is based on our proposed shifted diffusion model, which achieves better image embedding generation from input text. Unlike the baseline diffusion model used in DALL-E 2, our method seamlessly encodes prior knowledge of the pre-trained CLIP model in its diffusion process by designing a new initialization distribution and a new transition step of the diffusion. Compared to the strong DALL-E 2 baseline, our method performs better in generating image embedding from the text in terms of both efficiency and effectiveness, resulting in better text-to-image generation. Extensive large-scale experiments are conducted and evaluated in terms of both quantitative measures and human evaluation, indicating a stronger generation ability of our method compared to existing ones. Furthermore, our model enables semi-supervised and language-free training for text-to-image generation, where only part or none of the images in the training dataset have an associated caption. Trained with only 1.7% *of the images being captioned, our semi-supervised model obtains FID results comparable to DALL-E 2 on zero-shot text-to-image generation evaluated on MS-COCO. Corgi also achieves new state-of-the-art results across different datasets on downstream language-free text-to-image generation tasks, outperforming the previous method, Lafite, by a large margin.*

1. Introduction

"AI-generated content" has attracted increasingly more public awareness thanks to the significant progress in recent research of high-fidelity text-aligned image synthetic tasks. [7, 20–23, 26, 32]. Particularly, models trained on web-scale datasets have demonstrated their impressive ability to generate out-of-distribution images from arbitrary text inputs that describe unseen combinations of visual concepts.

Starting from DALL-E [21], researchers have proposed a variety of approaches to further advance the state-of-the-art (SOTA) of text-to-image generation in terms of both generation quality and efficiency. Latent Diffusion Model [22] trains a diffusion model in the latent space of auto-encoder instead of pixel space, leading to better generation efficiency. GLIDE [15] adopts a hierarchical architecture, which consists of diffusion models at different resolutions. Such a model design strategy has shown to be effective and has been adopted by many follow-up works. DALL-E 2 [20] further introduces an extra image embedding input. Such an image embedding not only improves the model performance in text-to-image generation but also enables applications, including image-to-image generation and generation under multi-modal conditions. Imagen [23] makes use of a rich pre-trained text encoder [19], demonstrating that a frozen text encoder pre-trained on the large-scale text-only dataset can help text-to-image generation models in understanding the semantics of text descriptions. Parti [32] shows a further successful scale-up of the generative model, leading to impressive improvement in text-to-image consistency with transformer structure.

The aforementioned approaches focus on improving text-to-image generation by either scaling up trainable/frozen modules or designing better model architectures. In this work, we explore an orthogonal direction, where we propose novel techniques to improve the diffusion process itself and make it more suitable and effective for text-to-image generation.

Specifically, we propose Corgi (**C**lip-based shifted diffusi**O**n model b**R**id**GI**ng the gap), a novel diffusion model designed for a flexible text-to-image generation. Our model can perform text-to-image generation under all the supervised, semi-supervised, and language-free settings. By "bridging the gap," we emphasize two key novelties in our method: (1) our model tries to bridge the image-text modality gap [12] so as to train a better generative model. Modality gap is a critical concept discovered in pre-trained models such as CLIP [17], which captures the phenomenon that multi-modality representations do not align in the joint em-

*Performed this work during internship at ByteDance, code is available at https://github.com/drboog/Shifted_Diffusion. The research of the first and last author was supported in part by NSF through grants IIS-1910492 and CCF-2200173 and by KAUST CRG10-4663.2.

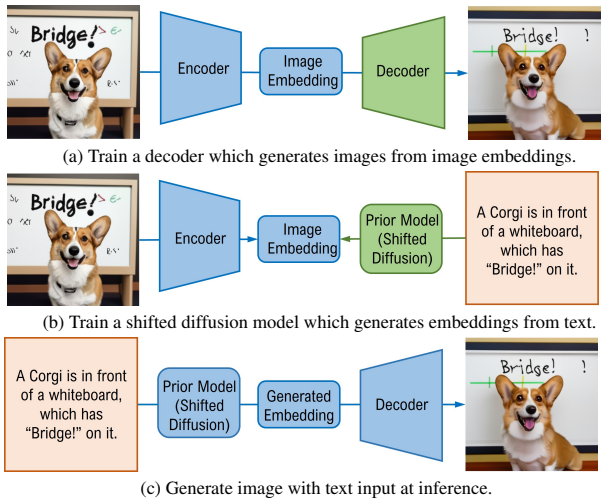
(a) A corgi in front of a blackboard with "Bridge!" on it. (b) A corgi with a neon light reads "Bridge!". (c) A realistic photo about corgi with word "Bridge!".

(d) A magazine about Corgi, which has "Bridge" on it. (e) A corgi is smiling, in front of a whiteboard with "Bridge the Gap" on it. (f) A corgi is giving a speech about "Bridge the Gap" in cartoon style.

Figure 1. We propose Corgi, a novel diffusion model designed for flexible text-to-image generation which can "bridge the gap".

bedding space. With Corgi, we can better utilize CLIP in text-to-image generation, which is shown in the experiment to achieve better generation; (2) our model bridges the gap of data availability for different researchers/communities. In our design, Corgi can naturally enable semi-supervised and language-free text-to-image generation, where only a small portion or even none of the images in the training dataset are captioned. This is important because the cost of constructing a high-quality large image-text-paired dataset could be prohibitive, especially in the case where hundreds of millions of images need to be captioned. We show that by merely using an image-only dataset and the public CC15M dataset [2,25], Corgi achieves promising results comparable to SOTA models on open-domain text-to-image generation tasks. To summarize, our contributions are:

- We propose Corgi, a novel diffusion model that seamlessly incorporates prior knowledge from the pre-trained model (*e.g.,*CLIP) into its diffusion process;
- Our method is general and can be applied in different settings of text-to-image generation, *e.g.*, it naturally enables semi-supervised and language-free text-to-image generation;
- Extensive and large-scale experiments are conducted. Both quantitative and qualitative results illustrate the effectiveness of the proposed method. We especially demonstrate that, with only 1.7% captioned images in the training dataset, it is possible to achieve results comparable to the SOTA method. In addition, our model achieves SOTA results under language-free setting across different downstream datasets.



(a) Train a decoder which generates images from image embeddings.

(b) Train a shifted diffusion model which generates embeddings from text.

(c) Generate image with text input at inference.

Figure 2. Illustration of our framework. The trainable modules are colored in green, and the frozen modules are colored in blue.

2. Methodology

We start by illustrating the proposed general framework for text-to-image generation. Our framework is shown in Figure 2, which consists of three key components: (1) a pre-trained image encoder that maps images to their embeddings; (2) a decoder that generates images from the corresponding embeddings; and (3) a prior model that generates image embeddings from the corresponding text captions. In our implementation, we use the pre-trained CLIP

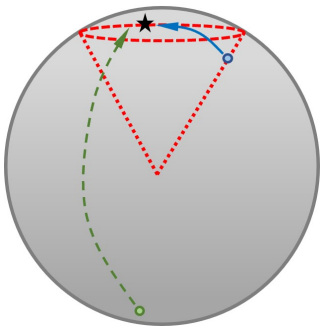


Figure 3. Illustration of diffusions inside the multi-modal joint space of CLIP. We use the cone in red dotted lines to represent the effective output space of CLIP image encoder. The target embedding is represented by the black star. The green dashed arrow represents the sampling process of the baseline diffusion model, whose starting point is random noise. The blue solid arrow represents the sampling process of our shifted diffusion, whose starting point is an image embedding inside the red cone.

image encoder because its output space is a multi-modal embedding space that has been demonstrated to benefit the text-to-image generation task [35]. The decoder can be either a diffusion model or a generative adversarial network (GAN). Note that if one chooses the decoder as a hierarchical diffusion model and makes it conditioned on both image embedding and text, our final structure will be similar to DALL-E 2 [20].

We adopt this framework because of its flexibility: it can perform different types of generation tasks such as text-to-image generation, image-to-image generation, and generation conditioned on both image and text. In addition, our framework naturally enables semi-supervised training, *i.e.*, the training dataset can be a mixture of image-text pairs and pure images that are not captioned. In this setting, the image-text pairs will be used to train the prior model, and all the pure images will be used to train the decoder. Such a semi-supervised training setting is important in practice, especially when training a text-to-image generation model on new domains with a limited budget. As discussed in [35], constructing high-quality image-text pairs could be very expensive and requires a heavy human workload. Our framework provides the community with the flexibility of choosing the number of images to be captioned based on their budget. As it will be shown in the experiments, semi-supervised training can obtain impressive results, which are even comparable to those of supervised training.

In this paper, we focus on improving the prior model, which is less exploited in previous works. As shown in DALL-E 2 [20], a diffusion-based prior model is introduced to generate the target CLIP image embedding $\mathbf{z}_0$ via the following sequential sampling process: $\mathbf{z}_{t-1} \sim p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{y})$ for $t = T, ..., 1$, where $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, t denotes timestep, $\mathbf{y}$ denotes text caption, and $p_{\boldsymbol{\theta}}(\cdot|\cdot)$ is the inverse transition distribution induced from the diffusion model. Although it is shown in [20] that this prior model can benefit generation in terms of both image-text alignment and image fidelity, we suspect that this vanilla sampling process may not be appropriate for generating high-quality CLIP image embeddings (which will be the inputs of decoder thus greatly influence the generation quality). The reason is that, as revealed in [12], the effective output space of the CLIP image encoder is actually a very small region of the whole embedding space, as shown in Figure 3. Consequently, $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which is the starting point of sampling, might be far away from the target embedding. Intuitively, if $\mathbf{z}_T$ is closer to the target $\mathbf{z}_0$, we may be able to well approximate $\mathbf{z}_0$ within fewer sampling steps. Likewise, we may better approximate the target within the same number of steps, if the initialization is closer to the target.

Based on this motivation, we propose shifted diffusion, a novel diffusion model that considers prior knowledge contained in the pre-trained CLIP image encoder. Specifically, the noise distribution $p(\mathbf{z}_T)$ of shifted diffusion is a parametric distribution rather than standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We present the details of our method in the following.

2.1. Shifted Diffusion

Let $q(\mathbf{z}_0)$ be the distribution of ground-truth image embeddings in the joint latent space of the CLIP model. Because of the fact that valid image embeddings only occupy a small region of the whole embedding space [12], in order to achieve better generation in a diffusion process, we would like to construct a new $p(\mathbf{z}^T)$ which is expected to be more similar to $q(\mathbf{z}^0)$ than standard Gaussian. However, $q(\mathbf{z}_0)$ is an intractable distribution, making it challenging to train a diffusion model based on it.

To tackle the aforementioned problem, we consider the initial distribution $p(\mathbf{z}_T)$ to be a parametric Gaussian distribution as $\mathcal{N}(\mathbf{z}_T; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and can be obtained by simply analyzing the training dataset*. We design the transition $q(\mathbf{z}_t \,|\, \mathbf{z}_{t-1})$ to be a Gaussian as

$$q(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}) := \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\,\mathbf{z}_{t-1} + \mathbf{s}_t, \beta_t \boldsymbol{\Sigma}), \qquad (1)$$

where β_t is a constant following [8]. Compared to the vanilla diffusion with

$$q(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}) := \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\,\mathbf{z}_{t-1}, \beta_t \mathbf{I}),$$

our diffusion process introduces a shift term $\mathbf{s}_t$ at every timestep t, thus termed *shifted diffusion*.

One can show that $q(\mathbf{z}_t \,|\, \mathbf{z}_0)$ has a closed-form expression (all proofs are provided in the Appendix):

$$q(\mathbf{z}_t \,|\, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sum_{i=1}^{t} \mathbf{s}_i \sqrt{\bar{\alpha}_t / \bar{\alpha}_i}, (1-\bar{\alpha}_t)\boldsymbol{\Sigma}),$$

*In practice, we set $\boldsymbol{\Sigma}$ to be a diagonal matrix with its element $\boldsymbol{\Sigma}_{i,i} = \kappa\sigma_i$. σ_i is the standard deviation of i^{th} element of all image embeddings from the training dataset, $\kappa > 0$ is a constant for scaling.

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. Specifically, we choose $\mathbf{s}_t = (1-\sqrt{1-\beta_t})\boldsymbol{\mu}$, leading to

$$q(\mathbf{z}_t \,|\, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + (1-\sqrt{\bar{\alpha}_t})\boldsymbol{\mu}, (1-\bar{\alpha}_t)\boldsymbol{\Sigma}). \quad (2)$$

Remark 1 *As we can see, by selecting $\{\beta_t\}_{t=1}^{T}$ such that $\alpha_T \approx 0$, $q(\mathbf{z}_T \,|\, \mathbf{z}_0)$ can approximate $\mathcal{N}(\mathbf{z}_T; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the distribution of the image embeddings. In other words, shifted diffusion process is a process that transfers a ground-truth image embedding into a random image embedding[†]; whereas the vanilla diffusion process is a process that transfers an image embedding into a random Gaussian noise, which is certainly not what we expect.*

From Equation (1) and (2), we can get the closed-form expression of $q(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}), q(\mathbf{z}_t \,|\, \mathbf{z}_0), q(\mathbf{z}_{t-1} \,|\, \mathbf{z}_0)$. By the property of Gaussian distribution [1] and some simple derivations, we can get the posterior distribution

$$\begin{aligned} &q(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\nu}, \boldsymbol{\Lambda}), \quad (3)\\ &\boldsymbol{\nu} = \gamma(\mathbf{z}_t - \mathbf{s}_t) + \eta\,\mathbf{z}_0 + \tau(1-\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\mu},\\ &\boldsymbol{\Lambda} = (1-\bar{\alpha}_{t-1})\beta_t\boldsymbol{\Sigma}/(1-\bar{\alpha}_t), \end{aligned}$$

where

$$\begin{aligned} \gamma &= (1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}/(1-\bar{\alpha}_t),\\ \eta &= \beta_t\sqrt{\bar{\alpha}_{t-1}}/(1-\bar{\alpha}_t),\\ \tau &= \beta_t/(1-\bar{\alpha}_t). \end{aligned}$$

With Equations (2) and (3), we can easily optimize the loss function of the proposed shifted diffusion

$$\begin{aligned} \mathbf{L}_{\boldsymbol{\theta}} = \mathbb{E}_q\{&\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_T \,|\, \mathbf{z}_0)\|p(\mathbf{z}_T)) - \log p_{\boldsymbol{\theta}}(\mathbf{z}_0 \,|\, \mathbf{z}_1) \quad (4)\\ &+ \sum_{t>1} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0)\|p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t))\}\,, \end{aligned}$$

where D_{KL} denotes KL-divergence. Since both $q(\mathbf{z}_t \,|\, \mathbf{z}_0)$ and $q(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0)$ are Gaussian distributions by our design, the KL-divergence terms have closed-form solutions, enabling easy stochastic optimization.

Although setting $p(\mathbf{z}_T)$ to be a Gaussian distribution leads to a simplified loss that can be conveniently optimized, it also introduces some drawbacks as the ground-truth distribution of image embedding is not a single-mode Gaussian. To tackle this problem, we propose to use a collection of Gaussian distributions, denoted as $\{p_i(\mathbf{z}_T)\}_{i=1}^{k}$ with $p_i(\mathbf{z}_T) := \mathcal{N}(\mathbf{z}_T; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Let $\mathbf{z}_0$ and $\mathbf{y}$ be the ground-truth image embedding and its associated text caption, respectively. For each pair of $(\mathbf{z}_0, \mathbf{y})$, we select its corresponding Gaussian $p_{c_\mathbf{y}}$ by the top-1 cosine similarity as

$$c_\mathbf{y} = \underset{1\le i\le k}{\operatorname{argmax}}\, \mathrm{Sim}(\boldsymbol{\mu}_i, f_{\text{txt}}(\mathbf{y})), \quad (5)$$

where $\boldsymbol{f}_{\text{txt}}$ is the pre-trained CLIP text encoder, and $c_\mathbf{y}$ denotes the index of the selected Gaussian. One may also use

$$c_\mathbf{y} = \underset{1\le i\le k}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{\epsilon}_i\sim p_i}\left[\mathrm{Sim}(\boldsymbol{\epsilon}_i, f_{\text{txt}}(\mathbf{y}))\right]\,, \quad (6)$$

which requires more computation because of the expectation. After selecting $p_{c_\mathbf{y}}(\mathbf{z}_T)$, its parameters $\boldsymbol{\mu}_{c_\mathbf{y}}$ and $\boldsymbol{\Sigma}_{c_\mathbf{y}}$ will be used in Equation (2) and (3) for optimization. An extra positional embedding representing $c_\mathbf{y}$ is also injected into the diffusion model, with a similar implementation as the time embedding for t. Compared to a single-mode Gaussian distributions, $\{p_i(\mathbf{z}_T)\}_{i=1}^{k}$ is supposed to have better expressive ability, and $p_{c_\mathbf{y}}$ is expected to better initialize $\mathbf{z}_T$ to make it closer to the target $\mathbf{z}_0$.

In our implementation, $\{p_i(\mathbf{z}_T)\}_{i=1}^{k}$ is estimated by performing clustering on the training dataset. Similar to existing quantization methods [28], we can also learn $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ by optimization. Specifically, we propose to update $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ during training with the following loss function

$$\begin{aligned} \mathbf{L}_p = &-\mathbb{E}_{\mathbf{z}_0}\{\mathbb{E}_{\mathbf{z}_T\sim p_{c_\mathbf{y}}(\mathbf{z}_T)}\{\mathrm{Sim}(\mathbf{z}_T, \mathbf{z}_0)\}\} \quad (7)\\ &+ \xi\frac{\sum_{i\neq j}}{k(k-1)}\mathbb{E}_{\mathbf{z}_T\sim p_i(\mathbf{z}_T)}\{\mathbb{E}_{\mathbf{z}'_T\sim p_j(\mathbf{z}_T)}\{\mathrm{Sim}(\mathbf{z}_T, \mathbf{z}'_T)\}\}, \end{aligned}$$

where ξ is a hyper-parameter. The first term of $\mathbf{L}_p$ forces the sample $\mathbf{z}_T$ from Gaussian $p_{c_\mathbf{y}}$ to be close to the corresponding ground-truth $\mathbf{z}_0$, and the second term ensures that $p_i(\mathbf{z}_T)$ does not overlap each other too much. Note that $\{p_i(\mathbf{z}_T)\}_{i=1}^{k}$ is only optimized with $\mathbf{L}_p$, *i.e.*, we manually stop the gradient back-propagation from $\mathrm{L}_{\boldsymbol{\theta}}$ to $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{k}$. Our algorithm is summarized in Algorithm 1, with more implementation details provided in the experiments section.

Algorithm 1 Shifted Diffusion

// Training
Require: $\{p_i(\mathbf{z}_T) = \mathcal{N}(\mathbf{z}_T; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i=1}^{k}$**, a diffusion model parameterized by $\boldsymbol{\theta}$, CLIP image encoder f_{img} and text encoder f_{txt}.**
while *not converge* **do**
  Sample image-text data pair $(\mathbf{x}_0, \mathbf{y})$, $\mathbf{z}_0 = f_{\text{img}}(\mathbf{x}_0)$
  Select corresponding $p_{c_\mathbf{y}}$ by (5)
  Update $\boldsymbol{\theta}$ by gradient descent w.r.t $\mathrm{L}_{\boldsymbol{\theta}}$
  if $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ are learn-able **then**
    Update $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ by gradient descent w.r.t L_p
  end if
end while
// Inference
Given a text caption $\mathbf{y}$, select its corresponding distribution $p_{c_\mathbf{y}}$, sample corresponding $\mathbf{z}_T$
for t=T,...,1 **do**
  Sample $\mathbf{z}_{t-1} \sim p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t)$
end for
Return $\mathbf{z}_0$

[†]The random image embedding might be scaled as we use a scaling factor κ.

(a) A photo of a light bulb in outer space traveling the galaxy with a sailing boat inside the light bulb.

(b) A high-resolution photo of a chicken working out in a gym.

(c) New York Skyline with 'Hello World' written with fireworks on the sky.

(d) A television made of water that displays an image of a cityscape at night.

(e) A photo of llama wearing sunglasses standing on the deck of a spaceship with the Earth in the background.

(f) A high contrast portrait photo of a fluffy hamster wearing an orange beanie and sunglasses holding a sign that says "Let's PAINT!".

(g) A photo of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.

(h) A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

Figure 4. Some generated examples on DrawBench and PartiPrompts.

3. Experiments

3.1. Zero-shot Text-to-image Generation

We first test the zero-shot text-to-image generation ability of Corgi. We prepare a dataset containing 900 million image-text pairs, which is composed of some commonly used datasets such as Conceptual Captions (CC3M) [25], Conceptual Captions 12M (CC12M) [2], filtered LAION-5B [24] and some image-text pairs collected by ourselves. Note that we make sure our dataset does not overlap with MS-COCO [13], CUB [29], Localized Narratives [16] and Multi-modal CelebA-HQ (MM-CelebA-HQ) [30], because we will test zero-shot or language-free performance on these downstream datasets.

Decoder Following [15, 20, 23], our decoder adopts a model architecture consisting of a hierarchy of diffusion models, which has shown impressive ability in text-to-image generation. Specifically, three diffusion models are trained on 64, 256, and 1024 resolutions, respectively. All the 900M images are used to train the decoder. During training, each image is processed by three different pre-trained CLIP models: ViT-B/16, ViT-B/32, and RN-101. The outputs from these three models will be concatenated into a single 1536-dimensional embedding, which is then projected into eight vectors and fed into the decoder (in the same manner as DALL-E 2).

Prior model To handle the image-text alignment, we train a shifted diffusion model to generate image embeddings from captions. Our shifted diffusion model is a decoder-only transformer whose input is a sequence consisting of encoded text from T5 [18], CLIP text embedding, an embedding representing diffusion timestep, an embedding representing the index of corresponding Gaussian, a noised CLIP image embedding and a final embedding which will be used to predict the target CLIP image embedding. We train two variants on datasets with different scales: one prior is trained on the full 900M image-text pairs; the other one is trained on CC15M[‡], which is a subset of our full dataset.

Final model Our two different prior models correspond to two final models. The decoders for the prior models (which are respectively trained on 900M and 15M image-text pairs) are trained on the 900M images. In other words, the two final models can be regarded as being trained in supervised and semi-supervised manners, where the training dataset of the first one consists of 900M image-text pairs, while the other one consists of 900M images but with only $1.7\% = 15/900$ of them associating with captions. We denote these two variants as Corgi and Corgi-Semi, respectively. More implementation details are provided in the Appendix.

In Table 1, we report zero-shot Fréchet Inception Distance (FID) evaluated on MS-COCO. Following previous

[‡]collection of CC3M and CC12M

(a) A panda making latte art. (b) Illustration of a mouse using a mushroom as an umbrella.

Figure 5. Comparison of our final models trained in supervised (left) and semi-supervised manner (right).

Methods	Supervised FID ↓	Zero-shot FID ↓
AttnGAN [31]	33.10	
DF-GAN [27]	21.42	
XMC-GAN [33]	9.33	
Lafite [35]	8.12	26.94
DALL-E [21]		27.50
CogView [4]		27.10
GLIDE [15]		12.24
LDM [22]		12.63
Make-A-Scene [6]		11.84
DALL-E 2 [20]		10.39
CogView 2 [5]		24.00
Imagen [23]		7.27
Parti [32]		7.23
Re-Imagen [3]		6.88
Corgi (Ours)		10.88
Corgi-Semi (Ours)		10.60

Table 1. Text-to-image generation results on MS-COCO.

works, the FID is calculated using 30,000 generated images, which corresponds to 30,000 randomly sampled captions from the validation set of MS-COCO. The results illustrate that our models obtain strong results, even when it is trained in a semi-supervised manner. Interestingly, our semi-supervised model obtains better FID than the supervised one. However, as we can see from Figure 5, the model trained with more image-text pairs leads to better image-text alignment and better image quality, as expected. One reason that it obtains relatively worse FID could be due to the dataset bias: CC15M might contain many samples that are similar to samples in MS-COCO, while this bias faded when they were merged to construct our full 900M dataset.

We note that most large-scale text-to-image generation models are trained on different datasets, but all tested on MS-COCO, ignoring potential dataset bias. Comparing these models simply by FID might not be appropriate. Thus, comparing models by actual generation quality is necessary. We follow previous works and choose to evaluate our model on DrawBench [23] and PartiPrompts [32]. Some generated examples with complicated scenes are shown in Figure 4. We compare our model with DALL-E 2 and Stable Diffusion, as these are the only models allowing public access. For a fair comparison, we fine-tune our model for an extra 300,000 iterations so that the decoder can take both image embeddings and text as inputs, which is similar to DALL-E 2. Some visual comparisons are provided in Figure 7. More results are provided in the Appendix. As can be seen, our model leads to better generation in most cases in terms of image-text alignment and image fidelity. Furthermore, we conduct a human evaluation on DrawBench and PartiPrompts. Specifically, we first generate four images for each prompt, then ten random human laborers are asked to judge which model is better (or comparable) in terms of image-text alignment and fidelity. The results are shown in Figure 6, where our model is shown to perform better in the two evaluation metrics consistently.

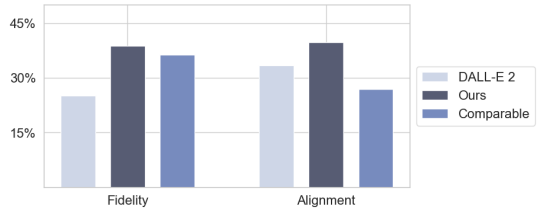


Figure 6. Human evaluation on DrawBench and PartiPrompts.

Methods	MS-COCO IS ↑	MS-COCO FID ↓	CUB IS ↑	CUB FID ↓	LN-COCO IS ↑	LN-COCO FID ↓
Lafite [35]	27.20	18.04	4.32	27.53	18.49	39.85
Lafite-2 [34]	31.16	10.26	4.93	16.87	23.18	25.51
Corgi (Ours)	**34.14**	10.33	**5.08**	**15.80**	**28.71**	**16.16**

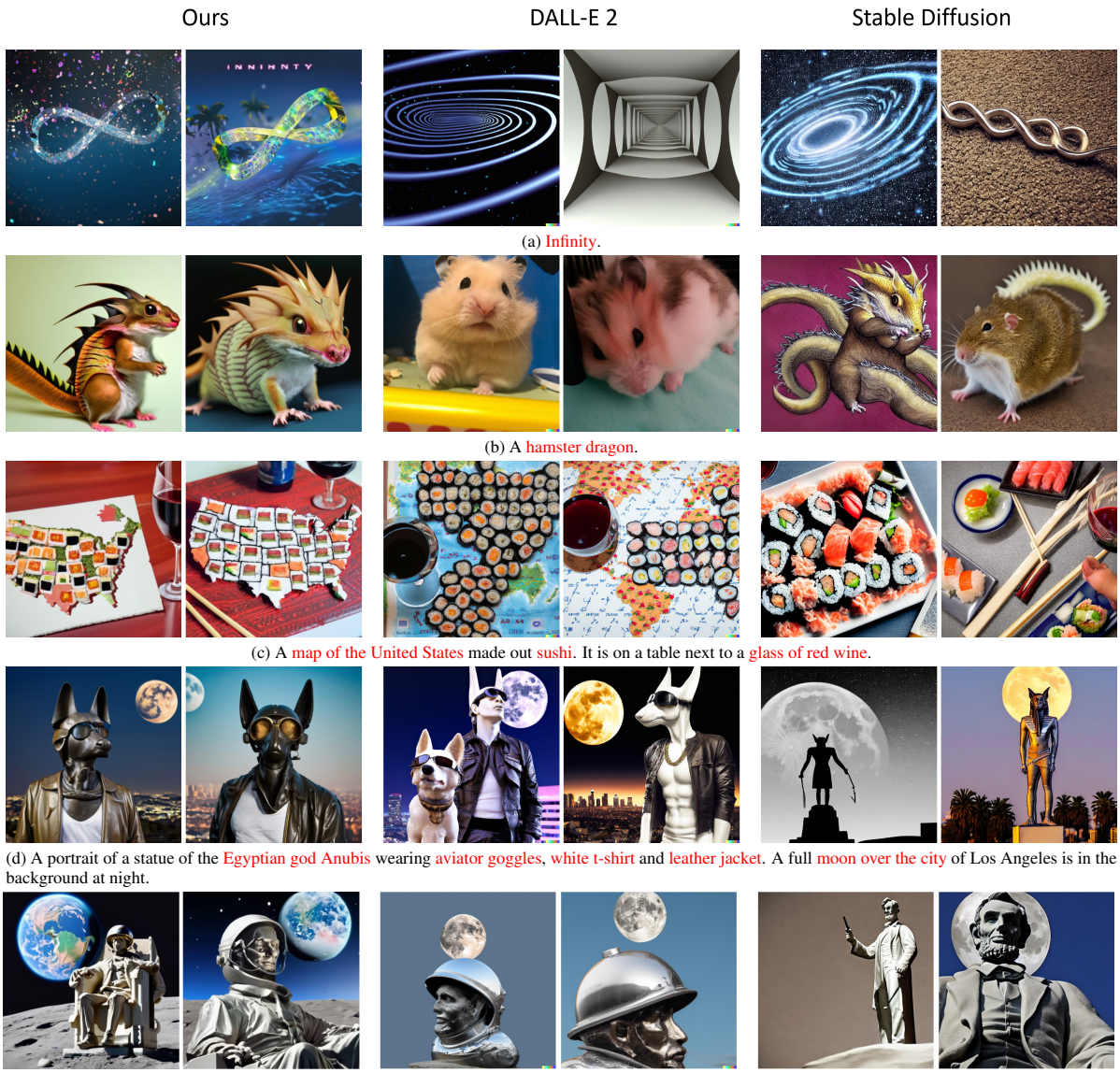
Table 2. Language-free results on different datasets.

3.2. Language-free Text-to-image Generation

[35] is the pioneering work to train text-to-image generation models on image-only datasets, termed language-free training as no associated captions are provided. With a pre-trained shifted diffusion model, we can also perform language-free training and fine-tuning on any downstream dataset. Given an image-only dataset, we can train or fine-tune a generative model that generates images from image embeddings. Consequently, we can directly use the pre-trained shifted diffusion model at inference time to perform text-to-image generation. Because our shifted diffusion model is pre-trained on large-scale image-text pairs, it is expected to generalize well on any downstream domain.

We first compare our method with [34, 35]. For a fair comparison, our decoder uses the same network architecture as [35], which is a StyleGAN2-based model. Some quantitative results are provided in Table 2, from which we can see that our method leads to better results in general. More results are provided in the Appendix. Although Lafite-2 [34] achieves a competitive FID on the MS-COCO dataset, it was trained on pseudo captions that require extra human workload. Specifically, to train Lafite-2, domain-specific vocabularies and prompts are needed, which requires human prior knowledge for each downstream domain. On the other hand, our shifted diffusion model is pre-trained and can be directly plugged into any domain without further training or fine-tuning.

We then conduct experiments of fine-tuning pre-trained text-to-image diffusion model under language-free setting. We choose Stable Diffusion 2 as our base model. Specifically, a projection layer is added to project CLIP image embedding into 4 word embeddings, which will be fed into the



(a) Infinity.

(b) A hamster dragon.

(c) A map of the United States made out sushi. It is on a table next to a glass of red wine.

(d) A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. A full moon over the city of Los Angeles is in the background at night.

(e) A statue of Abraham Lincoln wearing an opaque and shiny astronaut's helmet. The statue sits on the moon, with the planet Earth in the sky.

Figure 7. Comparison with DALL-E 2 and Stable Diffusion. More results are provided in the Appendix.

UNet of Stable Diffusion 2. As shown in Figure 8, Stable Diffusion 2 may generate images whose styles are different from target dataset, while our model leads to more satisfactory results after language-free fine-tuning.

3.3. Ablation Study

Baseline vs. shifted diffusion Although we have shown that our model obtains better generation quality than DALL-E 2 in previous experiments, there is no direct clue suggesting that the improvement is due to the advantages of shifted diffusion over the baseline diffusion. This is because, as we pointed out, other important factors, such as the potential dataset bias and different implementation details, may also affect model performance. Therefore, to better compare our shifted diffusion model with the baseline diffusion model, we conduct an ablation study where we train these two diffusion priors with the same implementation and training dataset (CC15M). We set $k = 1$ for shifted diffusion for a fair comparison, as a larger k leads to further improvements. More details are provided in the Appendix.

Firstly, we compare the cosine similarity of the generated image embeddings to the ground-truth image embeddings. We randomly sample 10,000 image-text pairs from the validation set of MS-COCO to prevent overlapping between training and testing datasets. The sampled captions are fed into different models to generate corresponding image embeddings. The results are shown in Figure 10. We can see that our shifted diffusion model leads to higher similarity scores than the baseline, implying better embedding generation.

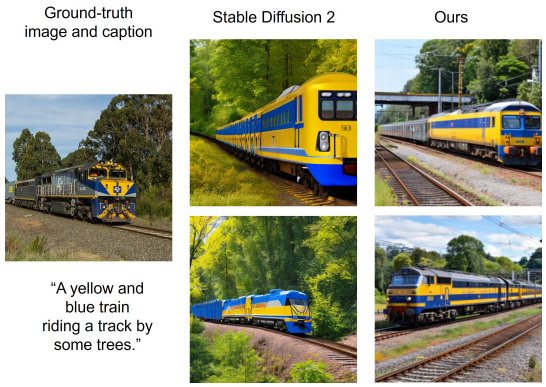
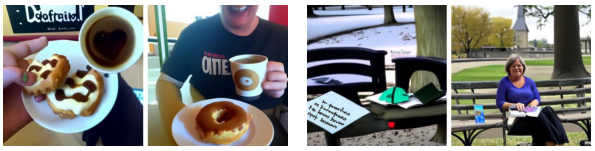


Figure 8. Pre-trained Stable Diffusion 2 vs. our fine-tuned model.



(a) A close-up photo of a person holding a cup and a doughnut

(b) A lady sitting on a park bench reading a book.

Figure 9. Generated examples with baseline diffusion (left) and shifted diffusion (right).

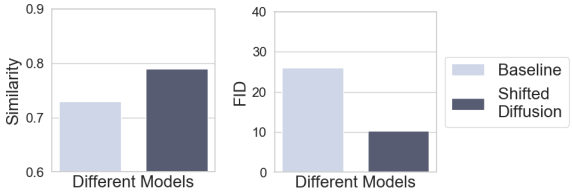


Figure 10. Comparison between baseline and shifted diffusion on our diffusion-based decoder which is trained from scratch.

Next, we compare the similarity scores of the generated embeddings to ground truth at different timesteps during the sampling process. We use 64 strided steps for both baseline and shifted diffusion model. The results are shown in Figure 12, from which we can see that our method leads to higher similarity scores, especially at initialization. This indicates that our starting point is much closer to the target, consistent with our design intention.

Finally, we compare the shifted diffusion and baseline by applying them in text-to-image generation tasks. To this end, we first generate 30,000 images using our diffusion-based decoder which is trained from scratch. The zero-shot FID results on MS-COCO are shown in Figure 10, along with some generated examples in Figure 9. We can see that shifted diffusion indeed leads to better quantitative and qualitative results, while the baseline model fails to capture some details in the text. We also evaluate them on fine-tuned Stable Diffusion 2, where FID and CLIP similarities of 10,000 generated images to ground-truth images and captions are calculated. The results are provided in Figure 11[§], from which we can find that shifted diffusion leads to better

[§]We report average similarity evaluated by ViT-B/16, ViT-B/32 and RN-101 CLIP models.

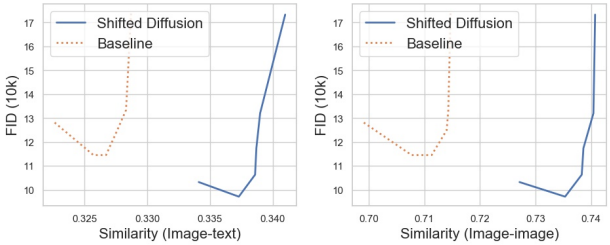


Figure 11. Comparison between baseline and shifted diffusion on fine-tuned Stable Diffusion 2.

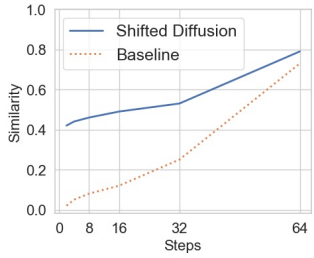


Figure 12. Evolution of embedding similarity during sampling.

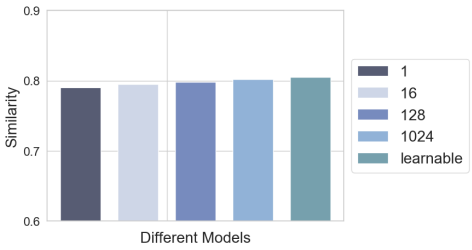


Figure 13. Results of shifted diffusion with different settings.

results as it obtains lower FID and higher CLIP similarities.

Different settings for shifted diffusion Recall that our shifted diffusion model adopts one or more Gaussian distributions as its initialization for the sampling process. We investigate the influence of the number of distributions in this experiment. We train our shifted diffusion model with 1, 16, 128, and 1024 Gaussians, respectively. Furthermore, we train a model with 1024 learn-able mean vectors and covariance matrices. All the models are trained on CC15M. We calculate the similarity of generated image embeddings to the ground-truth embeddings on the validation set of MS-COCO. The results are shown in Figure 13, from which we can see that using more Gaussian distributions leads to better results; furthermore, making the parameters learn-able gains further performance improvement. Some more discussions are provided in the Appendix.

4. Conclusion

We propose Corgi, a novel and general diffusion model that benefits text-to-image generation under different settings. Extensive large-scale experiments are conducted. Strong quantitative and qualitative results are obtained, illustrating the effectiveness of the proposed method.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. 4, 11, 12

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2, 5

[3] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 6

[4] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 6

[5] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 6

[6] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 6

[7] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 13

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 12

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

[12] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022. 1, 3

[13] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 13

[15] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 5, 6

[16] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer, 2020. 5

[17] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1

[18] 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. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 5

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 1

[20] 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, 3, 5, 6, 13

[21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 1, 6

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021. 1, 6

[23] 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, 5, 6

[24] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[25] 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, 5

[26] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 1

[27] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis, 2021. 6

[28] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning.

In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318, 2017. 4

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[30] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[31] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 6

[32] 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, 6

[33] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2021. 6

[34] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. Lafite2: Few-shot text-to-image generation. *ArXiv*, abs/2210.14124, 2022. 6, 12

[35] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021. 3, 6, 12, 13