# Towards Language-Free Training for Text-to-Image Generation

Yufan Zhou [1], Ruiyi Zhang [2], Changyou Chen [1], Chunyuan Li [3],
Chris Tensmeyer [2], Tong Yu [2], Jiuxiang Gu [2], Jinhui Xu [1]*, Tong Sun [2]
[1] State University of New York at Buffalo [2] Adobe Research [3] Microsoft Research, Redmond
{yufanzho, changyou, jinhui}@buffalo.edu
{ruizhang, tensmeye, tyu, jigu, tsun}@adobe.com chunyl@microsoft.com

2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-6946-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPR52688.2022.01738

## Abstract

*One of the major challenges in training text-to-image generation models is the need of a large number of high-quality image-text pairs. While image samples are often easily accessible, the associated text descriptions typically require careful human captioning, which is particularly time- and cost-consuming. In this paper, we propose the first work to train text-to-image generation models* without any text data*. Our method leverages the well-aligned multi-modal semantic space of the powerful pre-trained CLIP model: the requirement of text-conditioning is seamlessly alleviated via generating text features from image features. Extensive experiments are conducted to illustrate the effectiveness of the proposed method. We obtain state-of-the-art results in the standard text-to-image generation tasks. Importantly, the proposed language-free model outperforms most existing models trained with full image-text pairs. Furthermore, our method can be applied in fine-tuning pre-trained models, which saves both training time and cost in training text-to-image generation models. Our pre-trained model obtains competitive results in zero-shot text-to-image generation on the MS-COCO dataset, yet with around only 1% of the model size and training data size relative to the recently proposed large DALL-E model.*

## 1. Introduction

Automatic synthesis of realistic images from arbitrary text description is one of the core aspirations in artificial intelligence. Most existing works achieve the goal by consuming a large number of high quality image-text pairs [7,38,53,56,59], which, however, often requires heavy workload of precise human captioning and filtering. For instance, MS-COCO [27], the most commonly used dataset in text-to-image generation tasks, requires over 70,000 worker hours in gathering and annotating the captions. Even for less curated datasets such as Google Conceptual Captions [41], it consists of 3.3 million image-text pairs that are heavily filtered from 5 billion images from around 1 billion English webpages. In practice, for a customized domain, it is infeasible to collect such a large number of image-text pairs for model training, due to the high cost of human captioning and filtering. This challenge renders the unprecedented importance of the zero-shot text-to-image generation tasks, where no domain-specific image-text pairs are used to train a model to generate images in a given domain.

Recently, several attempts have been made to tackle zero-shot text-to-image generation problem, by pre-training giant generative models on web-scale image-text pairs, such as DALL-E [38] and CogView [7]. Both are auto-regressive Transformer models built for zero-shot text-to-image generation, as they can generate corresponding images given arbitrary text description without training on domain-specific datasets. However, to ensure good performance, these models require a gigantic scale of *data collections*, *model size* and *model training*. Specifically, DALL-E contains over 12 billion parameters and is trained on a dataset consisting of 250 million image-text pairs; CogView is a model with 4 billion parameters trained on 30 million image-text pairs. For this reason, hundreds of GPUs are required in training these models, which significantly increases carbon footprint and decrease the inclusivity: making it extremely difficult for more researchers to participate the study of this topic.

It is therefore desired to provide affordable solutions to build text-to-image generation models for the settings of limited image-text pair data, by reducing the requirements on model size, data collections and model training. In terms of data collections, in the ideal scenarios, the *language-free* setting is probably the minimal and cheapest requirement, where only image data is provided. This is important because collecting only image data is much easier than constructing high-quality image-text pairs, given the ample domain-specific image datasets available online.

*The research of the first and eighth authors was supported in part by NSF through grants IIS-1910492.

978-1-6654-6946-3/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPR52688.2022.01738

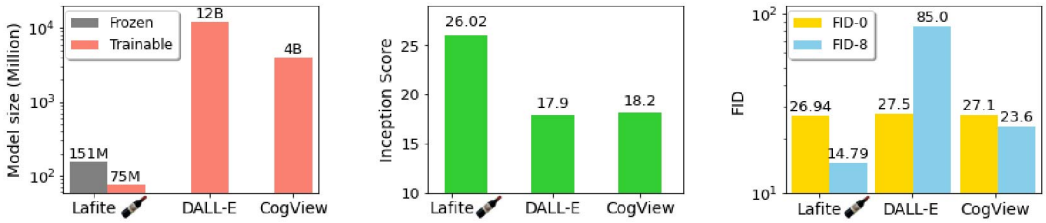



Figure 1. Model size *vs* performance of zero-shot image-to-text generation on the COCO dataset. LAFITE has much smaller model size, especially when considering trainable parameters (Left figure), but shows higher Inception score (Middle figure) and lower FID (Right figure). Please refer to Section 4 for details.

To this end, we propose LAFITE[1], a generative adversarial approach to significantly lowering the cost barrier and to building efficient text-to-image generation models, based on the pre-trained CLIP model [37]. Specifically, $(i)$ we take advantages of CLIP's property on image-text feature alignment in the joint semantic space, to construct pseudo image-text feature pairs; $(ii)$ we propose a text-to-image GAN (Generative Adversarial Network) model [11] that can effectively leverage pseudo image-text feature pairs. Our major contributions can be summarized as followings:

- We propose LAFITE, a versatile system that works effectively in a large range of text-to-image generation settings, including language-free, zero-shot and fully-supervised learning.
- To the best of our knowledge, LAFITE is the first work that enables the language-free training for the text-to-image generation task. We propose two novel schemes to construct pseudo image-text feature pairs, and conduct comprehensive study for the new setting. The effectiveness is validated with quantitative results on several datasets with different training schemes (training from scratch and fine-tuning from pre-trained generative models).
- In zero-shot text-to-image generation settings, LAFITE outperforms the prior art DALL-E and CogView on the COCO benchmark, with less than 1% of the trainable model parameter size (with frozen CLIP model weights). Please see Figure 1 for comparisons.
- In the standard fully supervised settings, LAFITE outperforms several state-of-the-art (SoTA) methods by a large margin. Surprisingly, even our language-free model shows superior performance than most existing models that are trained with full image-text pairs.

## 2. Related Work

**Text-to-image generation** Existing models on text-to-image generation can be categorized into two classes: fully-supervised text-to-image generation [53, 56, 59] and zero-shot text-to-image generation [7, 38]. The SoTA in the full image-text pair setting is still dominated by GAN variants [53,56,59]. GANs [11] have inspired many advances in image synthesis [18, 20, 23, 28, 32]. For text-to-image synthesis, the improved model performance is often benefited from large generative adversarial image models [56] and pre-trained text encoders [30]. Recently, excellent zero-shot text-to-image generation performance has been achieved in DALL-E [38] and CogView [7]. The basic idea is to encode images into discrete latent tokens using VQ-VAE [39, 45], and pre-train a huge-size auto-regressive Transformers [46] to predict these discrete tokens based on paired text sequences. Our LAFITE is the first generative adversarial approach that achieves SoTA on zero-shot generation.

**Multi-modal feature learning** Learning a joint and aligned feature space for vision-and-language has been a long standing problem in artificial intelligence [42, 50]. Inspired by the BERT model [6], a number of methods attempt to learn generic multi-modal fusion layers, given the pre-extracted visual region features and textual encoder [21, 24, 26, 31, 43, 57]. These works aim at learning generic multi-modal representations for downstream tasks like visual question answering [2, 14], image captioning [1, 27], visual commonsense reasoning [55]. Unlike the aforementioned works, another line of works focus on the way of learning visual representation from natural language supervisions, including both generative [5] and discriminative [48, 49, 58] methods. The latter learns an aligned visual-semantic space. This idea is recently scaled up in CLIP/ALIGN [16, 37], which pave the way toward building a *universal* image-text representation space. Our LAFITE is built up in this universal space, and is the first one to leverage its multi-modal alignment property for language-free text-to-image generation.

**CLIP for generation/manipulation.** The idea of multi-modal feature space also inspires some recent works on generative models [9, 10, 33, 35]. All of these works are related to ours in that the tools of pre-trained CLIP model

[1] **LA**nguage-**F**ree tra**I**ning for **T**ext-to-image g**E**neration

and StyleGAN2 are employed. Our LAFITE is different in two aspects: $(i)$ The motivations and scenarios are different. Existing works focus on latent optimization [10], image manipulation [35], domain adaptation [9], image segmentation [33]. We present the first study on training text-to-image generation models without the requirement of paired captions. $(ii)$ The techniques are different. Though image-text feature alignment property is leveraged in all works, Our LAFITE is the only one to generate pseudo features pairs in the joint multi-modal space, none of existing works considers such a possibility.

## 3. LAFITE: A Language-Free Paradigm

A natural idea to avoid human captioning in constructing image-text pair training data is using an off-the-shelf image captioning model that can automatically generate captions for the collected training images. However, this is especially challenging due to the lack of a universal captioning model that can $(i)$ bridge the modality gap between text and image to generate high-quality captions; $(ii)$ generalize to diverse image domains with large domain gaps. In this paper, we resort to solving an easier problem: one may directly *generate text features* rather than text descriptions, to avoid the use of image captioning models.

Throughout the paper, $(\mathbf{x}, \mathbf{t})$ denotes an image-text pair, $\mathbf{x}'$ is the corresponding generated image of $\mathbf{t}$. $G$ and $D$ denote the generator and discriminator respectively. We use $f_{\text{img}}$ and $f_{\text{txt}}$ to denote the pre-trained text encoder and image encoder, which map text descriptions and image samples into a joint multi-modal feature space. $\mathbf{h} = f_{\text{txt}}(\mathbf{t})$ denotes the real text feature, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ denotes latent noise sampled from the standard Gaussian distribution, serving as one input of the generator. Our idea to achieve language-free training is to generate pseudo text features $\mathbf{h}'$, which aims to approximating $\mathbf{h}$, by leveraging the image-text feature alignment of a pre-trained model. The generated features are then fed into the text-to-image generator to synthesize the corresponding images. Without loss of generality, we denote the mapping from input data to the multi-modal feature space as *translator* $T$ in two settings. If only images $\mathbf{x}$ are provided (*i.e.* language-free setting), we consider a pseudo text-feature generation process $T : \mathbf{x} \rightarrow \mathbf{h}'$; If image-text pairs $(\mathbf{x}, \mathbf{t})$ are provided (*i.e.* standard fully-supervised settings), we encode ground-truth text, $T : \mathbf{t} \rightarrow \mathbf{h}$.

### 3.1. Pseudo Text-Feature Generation

To achieve the goal, a universal multimodal feature space is desired, where features of paired texts and images are well aligned. The recently vision-and-language models such as CLIP and ALIGN achieve this, by pre-training on hundreds/thousands of millions of image-text pairs using contrastive learning. The cosine similarity between matched image-text features is maximized, while cosine similarity of the mis-matched pair is minimized. This naturally provides a high-dimensional hyper-sphere[2] for the multimodal features, where paired image-text should be close to each other, with a small angle between their feature vectors. This inspires us to explore the potentials of generating pseudo text features $\mathbf{h}' \in \mathcal{H}(\mathbf{x})$ for a given image $\mathbf{x}$ on this hyper-sphere: $\mathcal{H}(\mathbf{x}) = \{\mathbf{h}' | \text{Sim}(\mathbf{h}', f_{\text{img}}(\mathbf{x})) \geq c\}$, where Sim denotes cosine similarity, $c > 0$ is a threshold. This idea is illustrated in Figure 2. Based on the analysis, we consider two schemes to generate pseudo text features.

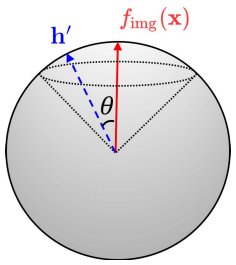



Figure 2. The illustration that the generated pseudo text feature vector $\mathbf{h}' \in \mathcal{H}(\mathbf{x})$ (blue dashed arrow) should have high cosine similarity with the image feature $f_{\text{img}}(\mathbf{x})$ (red solid arrow), *i.e.* $\theta \leq \arccos c$.

**Fixed perturbations** To generate pseudo text feature $\mathbf{h}'$, we propose to perturb the image feature $f_{\text{img}}(\mathbf{x})$ with adaptive Gaussian noise:

$$\mathbf{h}' = \tilde{\mathbf{h}}/\|\tilde{\mathbf{h}}\|_2, \quad \tilde{\mathbf{h}} = f_{\text{img}}(\mathbf{x}) + \xi\epsilon\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2, \quad (1)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian noise, $\xi > 0$ is a fixed hyper-parameter representing the level of perturbations, $\|\cdot\|_2$ denotes L2 norm. The added Gaussian noise is *adaptive* in the sense that it is normalized to a hyper-sphere, then re-scaled by the norm of image feature. We can prove that, with the adaptive noise, our LAFITE$_{\text{G}}$ can generate $\mathcal{H}(\mathbf{x})$ with a high probability which depends on $\xi, c$ and $d$. The formal theorem and its proof are provided in the Appendix.

**Trainable perturbations** It is natural to extend LAFITE$_{\text{G}}$ to learn more adaptive noise instead of using a vanilla Gaussian. To this end, we propose to train an *inference* model which takes the image features as inputs and outputs the mean and variance of the desired noise distribution. Specifically, the inference model consists of two neural networks $r_1(\cdot)$ and $r_2(\cdot)$. With the re-parameterization trick [22], the generation of pseudo text features is:

$$\mathbf{h}' = \tilde{\mathbf{h}}/\|\tilde{\mathbf{h}}\|_2, \text{where} \quad (2)$$
$$\tilde{\mathbf{h}} = f_{\text{img}}(\mathbf{x}) + r_1(f_{\text{img}}(\mathbf{x})) + \epsilon \odot \exp(r_2(f_{\text{img}}(\mathbf{x}))),$$

where exp denotes element-wise exponent operation, and $\odot$ denotes element-wise multiplication, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes noise sampled from standard Gaussian. In practice,

[2] In our implementation, we normalize the features extracted with CLIP by their L2 norm.

we construct $r_1(\cdot)$ and $r_2(\cdot)$ with 4 fully-connected (FC) layers respectively, and train them in a supervised way by maximizing the cosine similarity $\mathrm{Sim}(\mathbf{h}', \mathbf{h})$ between generated text features and real text features.

**Discussion.** Both schemes have their own pros and cons. The trainable perturbation generally yields better performance than the fixed perturbation. However, the fixed perturbation is easier to use, without the requirement of training an inference model on an additional dataset with annotated image-text pairs. Further, the performance of trainable perturbation is influenced by the gap between datasets used in training the inference model and the generative model, as empirically verified in our ablation studies in the experiment section.

### 3.2. Network Architectures

We propose to adapt the unconditional StyleGAN2 to a conditional generative model for our goal. Note that although we discuss our model in a language-free setting, it can be directly generalized to standard text-to-image generation by using $\mathbf{h}$ (real text feature) instead of $\mathbf{h}'$ (pseudo text feature).

**Generator** It is shown in recent works [29, 51] that the *StyleSpace* of StyleGAN2 is a well-disentangled intermediate feature space, whose dimensions are highly independent. By leveraging this property, we propose a simple yet effective approach to enable conditional generation: injecting new conditional information directly into the StyleSpace, as illustrated in Figure 3. Specifically, we choose to inject text information as follows. $(i)$ Random noise vectors $\mathbf{z} \in \mathcal{Z}$ are transformed into an intermediate latent space $\mathcal{W}$ via a so-called mapping network, which consists of a sequence of FC layers. The $\mathcal{W}$ space is claimed to better reflect the disentangled nature of the learned distribution. Each $\mathbf{w} \in \mathcal{W}$ is further transformed to channel-wise *unconditional style codes* $\mathbf{s}$, using a different learned affine transformation for each layer of the generator. The space spanned by these style parameters is often referred to as StyleSpace, or $\mathcal{S}$. $(ii)$ For a conditional vector $\mathbf{h}'$ from the image-text joint semantic space of CLIP, it is transformed into *condition codes* $\mathbf{c}$, using a different learned 2-layer FC network for each generator layer. $(iii)$ At each layer of the generator, we concatenate its style and conditional codes to obtain $[\mathbf{s}, \mathbf{c}]$, which is is further transformed to channel-wise *conditional style codes* $\mathbf{u}$ , using a different learned affine transformation for each generator layer. We refer to the space spanned by these style parameters as *Conditional StyleSpace*, or $\mathcal{U}$. In sum, the generator $G$ synthesizes a fake image as:

$$\mathbf{x}' = G(\mathbf{h}', \mathbf{z}) \tag{3}$$

**Discriminator** In the text-to-image task, the discriminator ensures the generated image to satisfy two criterias: photo-realistic to human perception and fidelity to the text condition. To this end, we encode the input image $\mathbf{x}$ with a shared discriminator backbone, then perform two tasks (each with a task-specific FC layer), as illustrated in Figure 4. $(i)$ $f_{\mathrm{d}}(\mathbf{x})$ projects $\mathbf{x}$ into a scalar, indicating the level of true or fake of an input image $\mathbf{x}$. This is a common task shared in all GAN models; $(ii)$ $f_{\mathrm{s}}(\mathbf{x})$ embeds $\mathbf{x}$ into a semantic space, which is expected to be similar to the semantic space of CLIP. We compute the inner product $\langle \mathbf{h}', f_{\mathrm{s}}(\mathbf{x}) \rangle$ to indicate how well the input image $\mathbf{x}$ is semantically aligned/conditioned with the pseudo text feature. In summary, the discriminator output is defined as:

$$D(\mathbf{x}, \mathbf{h}') = \underbrace{f_{\mathrm{d}}(\mathbf{x})}_{\text{real or fake}} + \underbrace{\langle \mathbf{h}', f_{\mathrm{s}}(\mathbf{x}) \rangle}_{\text{semantic alignment}}, \tag{4}$$

Intuitively, $D(\mathbf{x}, \mathbf{h}')$ yields a high value for an image $\mathbf{x}$, when it is real (with large $f_{\mathrm{d}}(\mathbf{x})$ values) and the semantic similarity between $\mathbf{h}'$ and $f_{\mathrm{s}}(\mathbf{x})$ is high. Similar ideas have been exploited in [15, 17, 56]. Different from these methods, our model can utilize the pre-trained multi-modal feature space, which relieves the difficulty for discriminator in learning semantically meaningful features.

### 3.3. Training Objectives

For a mini-batch of $n$ images $\{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{h}'_i$ is the corresponding generated pseudo text features for the $i$-th image. Our model is trained in an adversarial manner, with additional contrastive losses to ensure that the GAN feature space is aligned with pre-trained CLIP. The first one is the standard conditional GAN loss. The losses for the generator and discriminator are defined, with the logits from (4), as:

$$\mathcal{L}_{\mathrm{G}} = -\sum_{i=1}^{n} \log \sigma(D(\mathbf{x}'_i, \mathbf{h}'_i)), \tag{5}$$

$$\mathcal{L}_{\mathrm{D}} = -\sum_{i=1}^{n} \log \sigma(D(\mathbf{x}_i, \mathbf{h}'_i)) - \sum_{i=1}^{n} \log(1 - \sigma(D(\mathbf{x}'_i, \mathbf{h}'_i)))$$

where $\sigma(\cdot)$ denotes the Sigmoid function.

To enforce that the discriminator-extracted feature $f_{\mathrm{s}}(\mathbf{x})$ is semantically aligned in the pre-trained CLIP feature space, we consider the following contrastive regularizer for the discriminator:

$$\mathcal{L}_{\mathrm{ConD}} = -\tau \sum_{i=1}^{n} \log \frac{\exp(\mathrm{Sim}(f_{\mathrm{s}}(\mathbf{x}_i), \mathbf{h}'_i)/\tau)}{\sum_{j=1}^{n} \exp(\mathrm{Sim}(f_{\mathrm{s}}(\mathbf{x}_j), \mathbf{h}'_i)/\tau)}, \tag{6}$$

where $\mathrm{Sim}$ denotes the cosine similarity, $\tau$ is a non-negative hyper-parameter. Intuitively, $\mathcal{L}_{\mathrm{ConD}}$ enforces the discriminator to output image feature $f_{\mathrm{s}}(\mathbf{x}_i)$ that is similar to the corresponding text feature $\mathbf{h}'_i$.

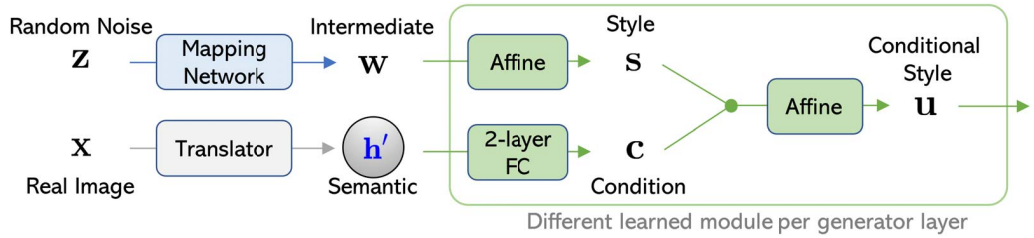
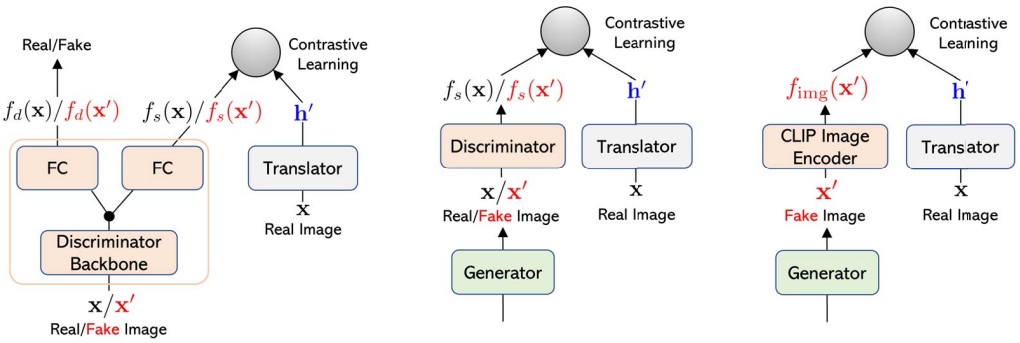



Figure 3. The process of injecting text-conditional information into each layer of the generator, where FC denotes fully-connected layer. The green modules have their own trainable parameters per generator layer. We can view the original StyleGAN2 constructs its StyleSpace as the process from $\mathbf{z}$ to $\mathbf{s}$. We propose to inject the semantic conditional information and further build our Conditional StyleSpace, whose elements $\mathbf{u}$ will be used to modulate image generation. This figure illustrates the language-free setting, where real image is used to generate pseudo text feature $\mathbf{h}'$; For the fully supervised text-to-image generation setting, real text is used for the extraction of text feature $\mathbf{h}$. Please refer to the definition of translator in Section 3 for details.




(a) Discriminator output (b) $\mathcal{L}_{\text{ConD}}$ (c) $\mathcal{L}_{\text{ConG}}$

Figure 4. Illustration of discriminator outputs and training objectives for the language-free setting.

We further utilize the pre-trained CLIP model to improve the semantic correspondence of the generated images $\mathbf{x}'_i$ and its conditioned pseudo text feature $\mathbf{h}'_i$. We define the following contrastive loss for the generator with the same hyper-parameter $\tau$ as (6):

$$\mathcal{L}_{\text{ConG}} = -\tau \sum_{i=1}^{n} \log \frac{\exp(\text{Sim}(f_{\text{img}}(\mathbf{x}'_i), \mathbf{h}'_i)/\tau)}{\sum_{j=1}^{n} \exp(\text{Sim}(f_{\text{img}}(\mathbf{x}'_j), \mathbf{h}'_i)/\tau)}. \quad (7)$$

With the above contrastive regularizers, the final training loss for the generator and discriminator are defined as:

$$\mathcal{L}'_{\text{D}} = \mathcal{L}_{\text{D}} + \gamma \mathcal{L}_{\text{ConD}} \quad (8)$$

$$\mathcal{L}'_{\text{G}} = \mathcal{L}_{\text{G}} + \gamma \mathcal{L}_{\text{ConD}} + \lambda \mathcal{L}_{\text{ConG}} \quad (9)$$

where $\tau = 0.5, \lambda = \gamma = 10$ for language-free settings, and $\tau = 0.5, \lambda = 10, \gamma = 5$ for fully-supervised settings[3].

### 3.4. Training Details

We summarize the language-free training schedule of LAFITE in Algorithm 1. For the settings with full image-text pairs, one may replace pseudo text feature generation step with the ground-truth text feature $\mathbf{h} = f_{\text{txt}}(\mathbf{t})$.

[3]Details about hyper-parameter tuning are provided in the Appendix.

**Algorithm 1** Language-free training of LAFITE

**Input:** An image dataset $\{\mathbf{x}_i\}_{i=1}^N$, pre-trained encoders $f_{\text{txt}}, f_{\text{img}}$, hyper-parameters $\tau > 0$
**while** *not converge* **do**
  Sample mini-batch $\{\mathbf{x}_i\}_{i=1}^n$;
  Sample perturbation noise $\{\epsilon_i\}_{i=1}^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  // Pseudo text feature generation
  Generate $\mathbf{h}'_i$ according to (1) or (2);
  // Forward pass of G and D
  Sample latent noise $\{\mathbf{z}_i\}_{i=1}^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  Synthesize fake image $\mathbf{x}'_i$ with G using (3);
  Feed real/fake images to D using (4);
  // Update G and D with gradient descent
  Update D with (8);
  Update G with (9);
**end while**

**Pre-training.** To demonstrate the zero-shot task transfer ability of our model, we also consider a variant that is pre-trained on the Google Conceptual Captions 3M (CC3M) dataset [41], which consists of 3.3 millions of image-text pairs. For pseudo text-feature generation with trainable per-

turbation, we also train its inference model on CC3M. There is no image overlapping between the pre-training and downstream datasets, which ensures the fairness when comparing our method against others in transfer learning. For face domain, we pre-trained a model on FFHQ dataset [19] which contains 70,000 images. The pre-trained models can be fine-tuned with LAFITE under language-free setting on different datasets, which will be discussed in next section.

**Data augmentation.** In practice, we also consider image data augmentation to improve extracted image features $f_{\text{img}}(\mathbf{x})$ in (1). We choose to use random cropping and avoid using augmentations like color transformation, because they may lead to mismatching between $\mathbf{h}'$ and $\mathbf{x}$. The details are summarized in Appendix.

## 4. Experiments

As the proposed LAFITE is a versatile system, we conduct experiments under different settings, including the proposed language-free setting, as well as the zero-shot and fully-supervised text-to-image generation settings. Due to the difference of two schemes to generate pseudo text features described in Section 3.1, we denote our system in two variants: fixed perturbations as LAFITE$_{\text{G}}$ and trainable perturbations as LAFITE$_{\text{NN}}$, respectively. All of our experiments are conducted on 4 Nvidia Tesla V100 GPUs, implemented using Pytorch [34]. CLIP-ViT/B-32 is used in our methods unless specified. All the codes and pre-trained models will be publicly available upon acceptance.

**Datasets.** We consider a suite of datasets that are commonly used in literature [53, 54, 56, 59], including MS-COCO [4], CUB [47], LN-COCO [36], Multi-modal CelebA-HQ (MM CelebA-HQ) [52]. All the images are scaled to resolution $256 \times 256$. Statistics of these datasets are summarized in Table 7 in the Appendix.

**Evaluation metrics.** Following [7, 38], we report the blurred Fréchet Inception Distance (FID) [12] and Inception Score (IS) [40] on MS-COCO dataset, which are computed using 30,000 generated images with randomly sampled text from validation set. FID-$k$ means the FID is computed after blurring all the images by a Gaussian filter with radius $k$.

### 4.1. Language-free Text-to-image Generation

We first study LAFITE under the proposed language-free setting, in which only images are provided in a given domain, and no paired caption is available during training.

**Captioning-based baseline:** As a baseline, we employed the SoTA image captioning model VinVL [57] to generate some associated captions for images. Note that MS-COCO image-text pairs were used to train the author-provided

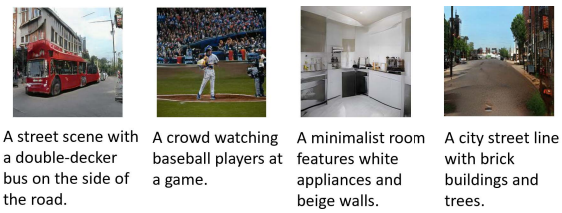



Figure 5. Language-free text-to-image generation examples on MS-COCO validation set.


Figure 6. Image generation with multi-modal conditions (conditioned on both image and text).

| Model | IS ↑ | FID-0 ↓ | FID-1 ↓ | FID-2 ↓ | FID-4 ↓ | FID-8 ↓ |
|---|---|---|---|---|---|---|
| Cap-Base | 15.83 | 56.36 | 54.99 | 51.84 | 44.81 | 37.28 |
| Cap-Large | 16.95 | 47.21 | 42.35 | 37.85 | 31.59 | 23.49 |
| LAFITE$_{\text{G}}$ | **27.20** | **18.04** | **17.80** | **17.68** | **16.16** | **14.52** |
| LAFITE$_{\text{NN}}$ | 22.23 | 26.56 | 26.48 | 25.82 | 23.90 | 19.27 |

Table 1. Results of language-free setting on MS-COCO dataset. 'Cap' indicates a text-to-image generation baseline method based on VinVL captioning.

VinVL image captioning model, so the MS-COCO comparison is unfairly biased in favor of the baseline due to this information leakage. We compare this baseline method with our LAFITE using the same network architecture and hyper-parameter setting for fairness. The main results are in Table 1. Both variants of our LAFITE significantly outperform the captioning-based baseline method. The simple LAFITE$_{\text{G}}$ performs the best on this dataset, indicating the generality of the method. For LAFITE$_{\text{NN}}$, note that CC3M is used to train the inference model, thus there is no information leakage in LAFITE$_{\text{NN}}$ method as we test LAFITE$_{\text{NN}}$ on the MS-COCO dataset. Some generated examples are provided in Figure 5, from which we can see that our LAFITE leads to text-aligned generation though no text data is used during training, verifying the effectiveness of the proposed method.

Furthermore, we can actually perform generation conditioned on images: For a given image, we generate an image-conditioned pseudo text feature vector with LAFITE. Passing this pseudo text feature vector to $G$ leads to generated

| Model | IS ↑ | FID-0 ↓ | FID-1 ↓ | FID-2 ↓ | FID-4 ↓ | FID-8 ↓ | SOA-C ↑ | SOA-I ↑ |
|---|---|---|---|---|---|---|---|---|
| DALL-E | 17.90 | 27.50 | 28.00 | 45.50 | 83.50 | 85.00 | - | - |
| CogView | 18.20 | 27.10 | 19.40 | 13.90 | 19.40 | 23.60 | - | - |
| LAFITE | **26.02** | **26.94** | 22.97 | 18.70 | **15.72** | **14.79** | 37.37 | 54.25 |

Table 2. Results of zero-shot setting on MS-COCO dataset, the model is pre-trained with image-text pairs from CC3M dataset.

| Model | MS-COCO | | | | CUB | | LN-COCO | | MM CelebA-HQ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | IS ↑ | FID ↓ | SOA-C ↑ | SOA-I ↑ | IS ↑ | FID ↓ | IS ↑ | FID ↓ | IS ↑ | FID ↓ |
| AttnGAN | 23.61 | 33.10 | 25.88 | 39.01 | 4.36 | 23.98 | 20.80 | 51.80 | - | 125.98 |
| Obj-GAN | 24.09 | 36.52 | 27.14 | 41.24 | - | - | - | - | - | - |
| DM-GAN | 32.32 | 27.34 | 33.44 | 48.03 | 4.75 | 16.09 | - | - | - | 131.05 |
| OP-GAN | 27.88 | 24.70 | 35.85 | 50.47 | - | - | - | - | - | - |
| DF-GAN | - | 21.42 | - | - | 5.10 | 14.81 | - | - | - | 137.60 |
| XMC-GAN | 30.45 | 9.33 | 50.94 | 71.33 | - | - | 28.37 | 14.12 | - | - |
| LAFITE | **32.34** | **8.12** | **61.09** | **74.78** | **5.97** | **10.48** | 26.32 | **11.78** | **2.93** | **12.54** |

Table 3. Standard text-to-image generation on CUB, LN-COCO and MM CelebA-HQ datasets.

images that are similar to the given image. Consequently, LAFITE enables image generation with multi-modal conditions, *i.e.* it can be conditioned on both image and text simultaneously. The implementation details are discussed in the Appendix. Some generated examples are provided in Figure 6, more results are provided in the Appendix.

### 4.2. Zero-Shot Text-to-image Generation

Zero-shot is a setting to evaluate a pre-trained text-to-image generation model, without training the model on any of downstream data. MS-COCO dataset is used for evaluating our model pre-trained on CC3M. The main results are shown in Table 2. Compared to DALL-E [38] and CogView [7], LAFITE achieves better quantitative results in most cases. We also emphasize that our model has only 75 millions of trainable parameters, while DALL-E has over 12 billions of parameters. Arguably, our pre-training dataset CC3M is much smaller[4], compared to the pre-training dataset used in DALL-E, which contains 250 millions of image-text pairs.

### 4.3. Standard Text-to-image Generation

We now consider the standard text-to-image generation task, where all the ground-truth image-text pairs are provided during training. We compare LAFITE against a series of competitive systems: AttnGAN [53], Obj-GAN [25], DM-GAN [59], OP-GAN [13], DF-GAN [44] and XMC-GAN [56]. The main results evaluated by FID and IS on different datasets are provided in Table 3. We also report the Semantic Object Accuracy (SOA) on MS-COCO following previous works [13, 56]. Results of competitive models are directly cited from the corresponding papers. It is clear that our proposed model consistently outperforms all other methods, creating new SoTA results in standard text-to-image generation.

[4]Though we acknowledge that LAFITE is based on an off-the-shelf discriminate model CLIP, which is trained on 400 million image-text pairs

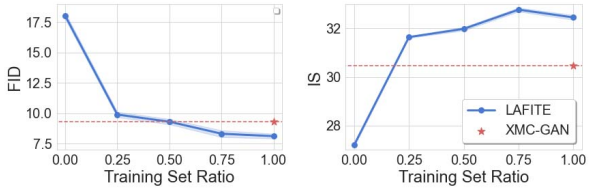
| Methods | MS-COCO | | CUB | | LN-COCO | | MM CelebA-HQ | |
|---|---|---|---|---|---|---|---|---|
| | IS ↑ | FID ↓ | IS ↑ | FID ↓ | IS ↑ | FID ↓ | IS ↑ | FID ↓ |
| | Training from Scratch | | | | | | | |
| LAFITE$_G$ | **27.20** | **18.04** | **4.32** | **27.53** | **18.49** | 38.95 | 2.78 | **32.75** |
| LAFITE$_{NN}$ | 22.23 | 26.56 | 4.06 | 46.32 | 18.17 | **36.19** | **2.89** | 50.34 |
| | Fine-tuned from Pre-trained Model | | | | | | | |
| LAFITE$_G$ | 24.89 | 20.89 | **6.13** | **35.99** | 19.32 | 34.96 | 3.10 | **15.74** |
| LAFITE$_{NN}$ | **26.55** | **17.44** | 4.36 | 37.91 | **20.02** | **33.76** | **3.19** | 29.42 |

Table 4. Comparisons between two schemes for language-free training on different datasets.

### 4.4. Adaptation of Pre-trained Models

**Language-free model fine-tuning.** Compared with existing works, one key advantage of the *pre-trained* LAFITE model is that it naturally enables language-free model fine-tuning. The results are provided in Table 4, where both LAFITE$_G$ and LAFITE$_{NN}$ are investigated on different datasets. We see that fine-tuning from the pre-trained model generally outperform training from scratch. We also notice that performance of pre-trained LAFITE largely depends on the domain gap in pre-training and fine-tuning datasets. For example, LAFITE$_{NN}$ sometimes obtains worse results than LAFITE$_G$, especially when the fine-tuning dataset is dissimilar to CC3M, *i.e.*, CUB and MM CelebA-HQ. This indicates that the inference model used for generating text features may have biases, because it may over-fit to its training dataset CC3M.

Pre-trained LAFITE is also highly training-efficient. For example, training from scratch with LAFITE on MS-COCO




(a) FID (↓) (b) IS (↑)

Figure 7. Comparison of LAFITE and prior art XMC-GAN. X-axis is the percentage of image-text pairs in the full MS-COCO dataset. XMC-GAN has over 166 millions trainable parameters, while our LAFITE only has 75 millions trainable parameters.

| Model | $\mathcal{L}_{\text{ConG}}$ | $\mathcal{L}_{\text{ConD}}$ | IS ↑ | FID ↓ | SOA-C ↑ | SOA-I ↑ |
|---|---|---|---|---|---|---|
| LAFITE$_G$ | | | 14.79 | 33.03 | 9.64 | 18.40 |
| | ✓ | | 17.78 | 29.65 | 16.53 | 30.33 |
| | | ✓ | 22.28 | 21.25 | 29.09 | 43.77 |
| | ✓ | ✓ | **27.20** | **18.04** | **36.84** | **54.16** |
| LAFITE$_{NN}$ | | | 11.05 | 72.03 | 8.28 | 14.46 |
| | ✓ | | 20.02 | 30.67 | 26.60 | 41.26 |
| | | ✓ | 19.14 | 33.88 | 33.32 | 49.86 |
| | ✓ | ✓ | **22.23** | **26.48** | **36.86** | **54.02** |

Table 5. Ablations of training losses on MS-COCO dataset, ✓ means the component is used during training.

dataset requires around 4 days to reach FID of 18, while fine-tuning only needs 3 hours. This becomes a critical advantage especially when we require several text-to-image generation models across different datasets.

**Semi-supervised fine-tuning.** Adaptation of pre-trained LAFITE is sample-efficient. One interesting question is, how much percentage of image-text pairs do we need to outperform previous SoTA XMC-GAN on MS-COCO dataset? To answer this question, we conduct experiment in which only a portion of the images are associated with ground-truth text. Our model is first pre-trained using all the images under the language-free setting, then it is fine-tuned with varying percentages of image-text pairs. The main results are summarized in Figure 7. Our method outperforms XMC-GAN on both IS and FID when less than half of total of the image-text pairs are employed.

### 4.5. Ablation Study

**Ablation study of training objectives** We first investigate the impact of each component in our objective functions. The standard generator and discriminator losses are always employed, we ablate by excluding $\mathcal{L}_{\text{ConG}}$ and $\mathcal{L}_{\text{ConD}}$ one by one. The results are provided in Table 5. For both variants of LAFITE, it is observed the model performance could drop significantly.

**Ablations of pre-trained text/image encoders** To demonstrate the importance of using a multi-modal feature-aligned pre-trained model in our LAFITE, we compare the CLIP model and other single-modality models. We adopt the popular RoBERTa [30] as the baseline text encoder, which was trained on a large text corpus only. Note that it is infeasible to perform language-free training without the joint feature space. Thus this experiment is based on fully-supervised text-to-image generation setting. For a fair comparison, we also report the results of only using the text encoder of CLIP while discarding the image encoder. In this setting, there is no image encoder thus the $\mathcal{L}_{\text{ConG}}$ term is removed from the objective function consequently. The results are reported in Table 6. As expected, even if the image encoder of CLIP is not used, models with only CLIP text encoder still significantly outperform models using RoBERTa. From the results, we can conclude that: $(i)$ The feature space of CLIP is semantically meaningful for text-to-image generation, thus only using text encoder of CLIP still leads to better results than RoBERTa; $(ii)$ Text-to-image generation results can be improved by using a feature-aligned joint feature space (CLIP vs others), and can be further improved with a stronger joint space (CLIP-ViT/B-16 outperforms CLIP-ViT/B-32, where ViT/B-16 and ViT/B-32 are different designs of visual transformers [8]).

| Model | Feature dim | IS ↑ | FID ↓ | SOA-C ↑ | SOA-I ↑ |
|---|---|---|---|---|---|
| RoBERTa-Base | 768 | 15.95 | 29.55 | 11.58 | 22.89 |
| RoBERTa-Large | 1024 | 14.11 | 35.77 | 7.72 | 16.03 |
| CLIP(B-32) Text encoder | 512 | 24.54 | 16.21 | 47.74 | 61.86 |
| CLIP(B-16) Text encoder | 512 | 24.90 | 15.97 | 47.80 | 62.71 |
| CLIP(B-32) | 512 | 31.88 | 8.62 | 59.51 | 73.76 |
| CLIP(B-16) | 512 | **32.34** | **8.12** | **61.09** | **74.78** |

Table 6. Results of using different pre-trained models on MS-COCO dataset.

## 5. Conclusion

We have presented LAFITE, an approach to build text-to-image generation systems without domain-specific image-text pairs in training. We achieve the goal by resorting to generating pseudo text features from images. Excellent performance in a variety of text-to-image generations tasks have demonstrated the effectiveness of LAFITE, including language-free, zero-shot and fully supervised settings. In particular, LAFITE creates new SoTA in zero-shot setting, with only 1% trainable parameter counts compared with recent advances such as DALL-E/CogView. LAFITE also outperforms prior arts in the fully-supervised settings. We believe that language-free training is a promising direction to enable broader application areas for text-to-image generation, as it significantly lowers the burden on data collection. One interesting future direction is to explore image synthesis in the wild, where long tail and open set conditions are provided for generation.

## References


[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Eungchun Cho. Inner product of random vectors. *International Journal of Pure and Applied Mathematics*, 56(2):217–221, 2009. 12

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014. 6

[5] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[7] 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. 1, 2, 6, 7

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 8

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2, 3

[10] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021. 2, 3

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[13] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 7

[14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[15] Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2020. 4

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 2

[17] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems*, 33:21357–21369, 2020. 4

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[19] 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. 6

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of styleGAN. *arXiv preprint arXiv:1912.04958*, 2019. 2

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[23] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. ALICE: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, 2017. 2

[24] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. 2

[25] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019. 7

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[27] 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. 1, 2

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 2

[29] Yunfan Liu, Qi Li, Zhenan Sun, and Tieniu Tan. Style intervention: How to achieve spatial disentanglement with style-based generators?, 2020. 4

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 8

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[33] Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E Green, and Nassir Navab. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. *arXiv preprint arXiv:2107.12518*, 2021. 2, 3

[34] 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. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 2, 3

[36] 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. 6

[37] 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. 2

[38] 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, 2, 6, 7

[39] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019. 2

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 6

[41] 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. 1, 5

[42] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973. IEEE, 2010. 2

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[44] 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. 7

[45] 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. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[47] 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. 6

[48] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 2

[49] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[50] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010. 2

[51] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 4

[52] 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. 6

[53] 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. 1, 2, 6, 7

[54] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021. 6

[55] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019. 2

[56] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2021. 1, 2, 4, 6, 7

[57] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2, 6

[58] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 2

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 1, 2, 6, 7