RLEG: Vision-Language Representation Learning with Diffusion-based Embedding Generation

Liming Zhao [1] **Kecheng Zheng** [2] **Yun Zheng** [1] **Deli Zhao** [1] **Jingren Zhou** [1]

Abstract

Vision-language representation learning models (*e.g.*, CLIP) have achieved state-of-the-art performance on various downstream tasks, which usually need large-scale training data to learn discriminative representation. Recent progress on generative diffusion models (*e.g.*, DALL-E 2) has demonstrated that diverse high-quality samples can be synthesized by randomly sampling from generative distribution. By virtue of generative capability in this paper, we propose a novel vision-language **R**epresentation **L**earning method with diffusion-based **E**mbedding **G**eneration (**RLEG**), which exploits diffusion models to generate feature embedding *online* for learning effective vision-language representation. Specifically, we first adopt image and text encoders to extract the corresponding embeddings. Secondly, pretrained diffusion-based embedding generators are harnessed to transfer the embedding modality online between vision and language domains. The embeddings generated from the generators are then served as augmented embedding-level samples, which are applied to contrastive learning with the variant of the CLIP framework. Experimental results show that the proposed method could learn effective representation and achieve state-of-the-art performance on various tasks including image classification, image-text retrieval, object detection, semantic segmentation, and text-conditional image generation.

1. Introduction

Vision-language representation learning (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021; Singh et al., 2022; Wang et al., 2022a; Yu et al., 2022) has received increasing attention and achieved remarkable success in pretraining for many computer vision tasks. Large-scale training data make it possible to transfer pretraining models to various downstream tasks. However, collecting a high-quality dataset for image-text pairs is non-trivial (Li et al., 2022b; Schuhmann et al., 2021; Byeon et al., 2022). Furthermore, the collected image-text pairs can only cover part of semantic contents from all possible real-world data, thus resulting in a sparse distribution of learned representation in the embedding space. In this paper, we attempt to learn robust representation by generating training samples of rich diversity online with generative models.

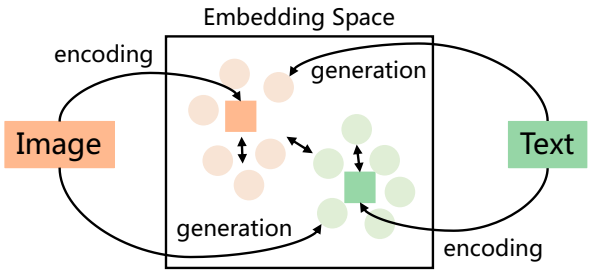

Figure 1. Input samples of image and text are encoded into a semantic embedding space (squares). More generated embeddings (circles) are utilized by sampling from pretrained diffusion-based embedding generators. The input and generated embedding samples are simultaneously applied to a contrastive learning scheme.

Recent advances in deep generative models (Brock et al., 2019; Karras et al., 2020; van den Oord et al., 2017; Ho et al., 2020; Dhariwal & Nichol, 2021; Ramesh et al., 2022; Saharia et al., 2022) make it possible to produce various high-quality realistic samples with different content under a semantic condition. Traditional image-level or feature-level data augmentation usually make fake modifications of input data (*e.g.*, adding noise, interpolating, random cropping), while sampling data from a learnt generative space could generate totally new and diverse realistic samples. Some methods (Bowles et al., 2018; Frid-Adar et al., 2018; Zhang et al., 2022; Liu et al., 2022) successfully augment training data from generative models to acquire synthetic medical images, foggy images, and images in different poses or viewpoints. These works show that the pretrained generative

[1]Alibaba Group [2]Ant Group. Correspondence to: Liming Zhao <lingchen.zlm@alibaba-inc.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

models can learn extremely powerful latent representations, which can be utilized for vision-language representation learning.

We are also inspired by a hypothesis in generative modeling and manifold learning stating that real-world data tend to lie on a low-dimensional manifold embedded in a high-dimensional space (Roweis & Saul, 2000; Song & Ermon, 2020). The process of vision-language representation learning is to update the model to produce embeddings distributed in such a manifold during training. In theory, generative models could provide a dense manifold matched with real-world data, and sampling data from such distribution is beneficial to learning better representation. In this paper, we train the model to match the embeddings sampled from generative models, which can also be regarded as distilling knowledge from a well-learnt dense manifold.

An important goal of vision-language representation learning is to align the embeddings of image and text pairs (Radford et al., 2021), which are usually collected from the Web. We argue that aligning an image with only one text sample is not enough since "a picture is worth a thousand words". Sampling data with similar semantic information from generative models will enrich the content of image and text for alignment beyond the training dataset. In this paper, multiple embedding samples are generated to enhance the original embedding matching process.

In principle, we exploit the diffusion models (Ho et al., 2020; Ramesh et al., 2022) in latent embedding space to help train a better foundation model for vision-language representation learning. Specifically, we first encode the input image and text by an image encoder and a text encoder respectively to obtain *input embeddings* for alignment. Secondly, diffusion-based embedding generators are applied for sampling *generated image embeddings* transferred from the *text input embedding* and sampling *generated text embeddings* transferred from the *image input embedding*, as shown in Figure 1. Multiple samplings are used to produce more *generated embeddings* for effective data augmentation in the feature space. At last, we align the embeddings between input embeddings and generated embeddings simultaneously in a unified contrastive learning scheme. Extensive experiments are conducted to analyze the proposed framework, and results on various downstream tasks demonstrate the effectiveness of our method on vision-language representation learning.

To summarize, the main contributions of this work are listed as follows.

- We present a novel framework for learning effective vision-language representation using diffusion-based embedding generators.
- We successfully integrate generative models into contrastive learning models with cross-modality embedding generation.
- We evaluate the effectiveness of our method on various tasks including image classification, image-text retrieval, object detection/segmentation, and text-conditional image generation.

2. Related works

2.1. Vision-Language Representation Learning.

Learning representation from vision-language data has received much attention and achieved significant progress in the computer vision community. There are two main streams of vision-language learning frameworks, *i.e.*, contrastive-learning frameworks, and generative-learning frameworks. Contrastive-learning frameworks (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021; Li et al., 2022b) demonstrate that using language text as supervision is able to obtain a good image encoder, which makes the pretraining scalable since image-text pairs is easier to access than label annotations. Contrastive loss is used in these methods to minimize the distance between embeddings of image-text pairs. Pioneering methods such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) introduce large-scale dataset to learn transferable visual features, which achieve impressive success on different downstream vision tasks, including image classification and cross-modal retrieval in zero-shot settings.

Generative-learning frameworks (Yu et al., 2022; Wang et al., 2022a;b; Dong et al., 2023) introduce the image or text generation task beyond contrastive learning with image-text pairs. The CoCa method (Yu et al., 2022) combines image captioning and contrastive learning to align image and text embeddings. GIT (Wang et al., 2022a) boosts the pretraining with language modeling to predict image caption with a concatenation of image and text tokens. Currently, some methods (Wang et al., 2022b; Dong et al., 2023) exploit masked image modeling to further improve the performance of the visual encoder. The generative tasks used in representation learning enable the improvement of various multimodal downstream tasks, such as image captioning (Lin et al., 2014) and visual question answering (VQA) (Zhou et al., 2020). In this paper, we utilize generative models for generating image and text embeddings online to improve the quality of both feature encoding and image-text alignment.

2.2. Feature-level Augmentation.

The common method of feature-level data augmentation is to sample new features from a hypothesized data distribution while keeping the semantic concepts not changed. Most methods model the data distribution by modifying the

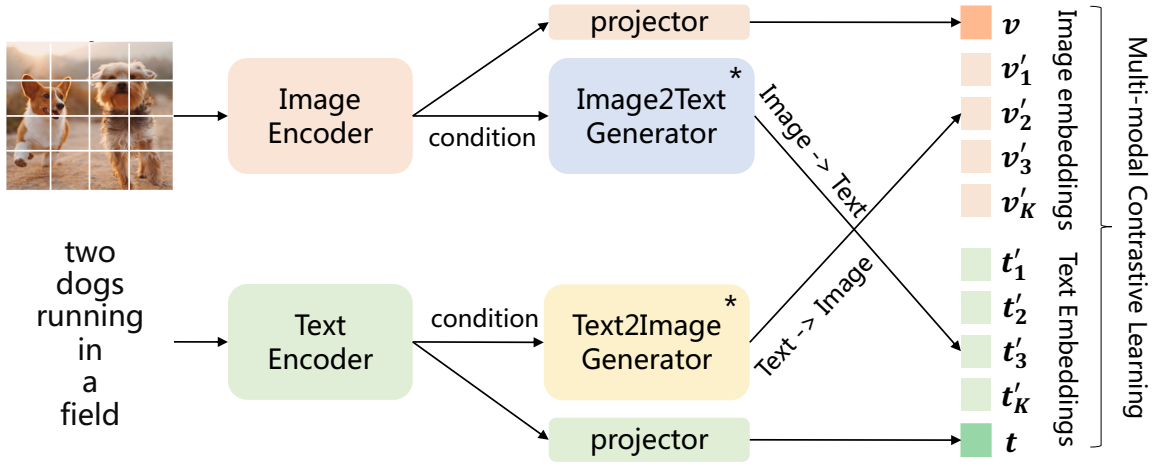

Figure 2. Framework of our **RLEG**. We first use image and text encoders to extract the corresponding embeddings. Then, the pretrained diffusion-based embedding generators are used to transfer the embedding modality online between vision and language domains. The generated embeddings are served as augmented embedding-level samples, which are then applied to contrastive learning.

original input data.

Li et al. (Li et al., 2022a) aim to sample features from a hypothesized multivariate Gaussian distribution. The authors in (Kumar et al., 2019) propose a Linear Delta method by simply adding the difference between two examples from the same class to a new example. Dai et al. (Dai et al., 2019) randomly erase feature elements with dropout and then combine them with original feature. A more detailed survey about image-level and feature-level data augmentation can be found in (Mumuni & Mumuni, 2022). In contrast to modifying the original input data, the proposed method exploits diffusion models to generate totally new samples from a well-learnt generative distribution.

2.3. Diffusion-based Generative Models.

Diffusion-based generative models (Sohl-Dickstein et al., 2015; Mehrjou et al., 2017; Sajjadi et al., 2018; Song et al., 2021; Ho et al., 2020; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Ho et al., 2022a; Ramesh et al., 2022; Saharia et al., 2022) have witnessed significant advances in recent years. The generating targets are various, such as image synthesis (Sohl-Dickstein et al., 2015), video prediction (Ho et al., 2022b; Höppe et al., 2022), 3D shapes (Cai et al., 2020), and graph generation (Niu et al., 2020).

Diffusion probabilistic models (Sohl-Dickstein et al., 2015; Ho et al., 2020) define a process of denoising from the Gaussian distribution step-by-step to generate samples from the real-world data distribution. In such training framework, various architectures are proposed and have achieved impressive generation results (Dhariwal & Nichol, 2021; Sajjadi et al., 2018; Ho et al., 2022a; Mehrjou et al., 2017). In the setting of vision-language tasks, text-to-image diffusion models (Ramesh et al., 2022; Saharia et al., 2022) receive much attention and become popular in the community. Both DALL-E 2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022) utilize diffusion models to generate images from text embeddings. Imagen (Saharia et al., 2022) directly learns the model from text-embedding inputs. DALL-E 2 (Ramesh et al., 2022) first learns a diffusion decoder from the image-embedding input, and a prior diffusion model is proposed for translating the text embedding to a corresponding image embedding, which enables the text-conditioned image generation. In this paper, the diffusion models of dually translating image-text embeddings are adopted to generate samples online for the purpose of representation learning.

3. Method

The proposed framework consists of two main components, as shown in Figure 2. One is image and text encoders to extract the corresponding embeddings. The second is pretrained diffusion-based embedding generators to transfer the embedding modality online between vision and language domains. Then, unified losses are applied on the embeddings to learn vision-language representation.

3.1. Vision-Language Contrastive Learning

We first revisit vision-language models based on contrastive learning. Given a set of image-text pairs $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$, where $\boldsymbol{x}_i$ is an input image and $\boldsymbol{y}_i$ is its corresponding text description, an image encoder (*e.g.*, ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2021)) is adopted to extract image feature vectors $\{\boldsymbol{v}_i\}_{i=1}^N$, and a text encoder (*e.g.*, BERT (Devlin et al., 2019)) is used to extract text feature vectors $\{\boldsymbol{t}_i\}_{i=1}^N$. The feature vectors $\boldsymbol{v}_i$ and $\boldsymbol{t}_i$ are of the same di-

mension and L2-normalized. During training, contrastive loss is applied on a batch of image-text embedding pairs $\{\boldsymbol{v}_i, \boldsymbol{t}_i\}_{i=1}^{B}$, which aims to reduce the distances of matched image-text pairs while enlarging the distances between unmatched ones.

In this paper, following the standard method CLIP (Radford et al., 2021) and the unified contrastive learning methods (Yang et al., 2022; Yuan et al., 2021), we adopt the extension of transforming image-text pairs from one-to-one mapping to one-to-multiple mapping relation. Specifically, an input image $\boldsymbol{x}_i$ is supposed to have only one corresponding description text $\boldsymbol{y}_i$ in one-to-one mapping assumption. Under the one-to-multiple relation, there are a set of text embeddings $\{\boldsymbol{t}_k\} \in \mathcal{R}(i)$ corresponding to the i-th image embedding $\boldsymbol{v}_i$ in a large batch. The set $\mathcal{R}(i)$ of *image to multiple texts* is pre-computed on the input dataset such that the text descriptions share the same valid words after text normalization (*e.g.*, case normalization and text cleaning by removing unrecognizable symbols, redundant spaces, and so on), and the image duplication is also discovered to compute the one-to-multiple relation set of *text to multiple images*.

Formally, we use the modified InfoNCE (van den Oord et al., 2018) loss upon a batch of image-text embeddings. The image-to-text contrastive loss is to align the image embedding $\boldsymbol{v}_i$ with a set of text embeddings $\mathcal{R}(i)$

$$\mathcal{L}_{i2t} = -\sum_{i=1}^{B} \frac{1}{|\mathcal{R}(i)|} \sum_{r \in \mathcal{R}(i)} \log \frac{\exp(\boldsymbol{v}_i^T \boldsymbol{t}_r / \tau)}{\sum_{j=1}^{N} \exp(\boldsymbol{v}_i^T \boldsymbol{t}_j / \tau)}, \quad (1)$$

where τ is a learnable temperature parameter to scale the pairwise cosine similarities. The text-to-image contrastive loss is to align the text embedding $\boldsymbol{t}_j$ with a set of image embeddings $\mathcal{R}(j)$

$$\mathcal{L}_{t2i} = -\sum_{j=1}^{B} \frac{1}{|\mathcal{R}(j)|} \sum_{r \in \mathcal{R}(j)} \log \frac{\exp(\boldsymbol{v}_r^T \boldsymbol{t}_j / \tau)}{\sum_{i=1}^{N} \exp(\boldsymbol{v}_i^T \boldsymbol{t}_j / \tau)}. \quad (2)$$

In Equations (1) and (2), it would be the original InfoNCE loss used in CLIP if there is only one relevant data point for each i-th sample (*i.e.*, the set size $|\mathcal{R}(i)| = |\mathcal{R}(j)| = 1$).

3.2. Diffusion-based Embedding Generation

Diffusion-based embedding generators are used as pretrained generative models in the proposed framework to translate embeddings between image and text domains. We follow the "prior" models used in DALL-E 2 (Ramesh et al., 2022) to obtain the diffusion models, which are originally proposed to translate the CLIP text embedding to corresponding image embedding for the purpose of text-conditioned image generation.

Preliminary. To introduce the process of embedding generation, we give a brief overview of the diffusion-based generative models. Given an image embedding $\boldsymbol{v}_0 \in q(\boldsymbol{v}_0)$, the diffusion process gradually adds Gaussian noise with variance $\beta_t \in (0, 1)$ at time t in total T steps, which produces a sequence of latent variables $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$. Formally, the diffusion process or forward process q is defined as follows:

$$q(\boldsymbol{v}_{1:T}|\boldsymbol{v}_0) \coloneqq \prod_{t=1}^{T} q(\boldsymbol{v}_t|\boldsymbol{v}_{t-1}), \quad (3)$$

$$q(\boldsymbol{v}_t|\boldsymbol{v}_{t-1}) \coloneqq \mathcal{N}(\boldsymbol{v}_t; \sqrt{1-\beta_t}\boldsymbol{v}_{t-1}, \beta_t \mathbf{I}), \quad (4)$$

where $\mathcal{N}$ is a Gaussian distribution. By defining $\alpha_t \coloneqq 1 - \beta_t$ and $\bar{\alpha}_t \coloneqq \prod_{i=0}^{t} \alpha_i$, we can rewrite Equation (4) as:

$$q(\boldsymbol{v}_t|\boldsymbol{v}_0) = \mathcal{N}(\boldsymbol{v}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{v}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (5)$$

$$\boldsymbol{v}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{v}_0 + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\epsilon}, \quad (6)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is a random noise. The final $\boldsymbol{v}_T$ at a large time T will be nearly a pure noise in $\mathcal{N}(0, \mathbf{I})$.

Usually, a reverse process $q(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t)$ is performed for generating samples by diffusion models. Starting from sampling $\boldsymbol{v}_T$ from $\mathcal{N}(0, \mathbf{I})$, we approximate the reverse process with $p_\theta(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t)$ to produce $\boldsymbol{v}_0$ step-by-step:

$$p_\theta(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t) \coloneqq \mathcal{N}(\boldsymbol{v}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{v}_t), \sigma_t^2 \mathbf{I}), \quad (7)$$

where σ_t is a variance constant (Ho et al., 2020), and $\boldsymbol{\mu}_\theta(\boldsymbol{v}_t)$ could be estimated by a deep model with learnable parameters θ.

Given an input sample $\boldsymbol{v}_0$, we can calculate the posterior $q(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t, \boldsymbol{v}_0)$ using Bayes theorem (Ho et al., 2020; Nichol & Dhariwal, 2021):

$$q(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t, \boldsymbol{v}_0) = \mathcal{N}(\boldsymbol{v}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{v}_t, \boldsymbol{v}_0), \tilde{\beta}_t \mathbf{I}), \quad (8)$$

where $\tilde{\boldsymbol{\mu}}_t(\boldsymbol{v}_t, \boldsymbol{v}_0)$ and $\tilde{\beta}_t$ are:

$$\tilde{\beta}_t \coloneqq \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \quad (9)$$

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{v}_t, \boldsymbol{v}_0) \coloneqq \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{v}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{v}_t. \quad (10)$$

In the training phase, a noisy $\boldsymbol{v}_t$ could be sampled by Equation (6) on input $\boldsymbol{v}_0$, and we can train the diffusion model $\boldsymbol{\mu}_\theta$ to directly predict $\tilde{\boldsymbol{\mu}}_t$ computed by Equation (10). In practice (Ho et al., 2020), the network could predict the noise $\boldsymbol{\epsilon}$ with a re-weighted loss function $\mathcal{L}_{\text{simple}} = \mathbb{E}_{\boldsymbol{v}_0, \boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t)||^2\right]$ derived by Equation (6) and Equation (10).

Generating Process. Given a pretrained deep diffusion generative model $\boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t, \boldsymbol{t})$ conditioned on text embedding $\boldsymbol{t}$,

we could sample data $\boldsymbol{v}_0$ starting from a normal distribution $\boldsymbol{v}_T \sim \mathcal{N}(0, \mathbf{I})$ step-by-step:

$$\boldsymbol{v}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{v}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t, \boldsymbol{t}) \right) + \sigma_t \boldsymbol{\epsilon}. \quad (11)$$

Classifier-free guidance (Ho & Salimans, 2021) is used in the embedding generators to enable the translation between image and text embeddings. Specifically, when generating the image embedding from a normal distribution, the text embedding $\boldsymbol{t}$ is added as a condition to guide the sampling process, and the generation is dependent on both the unconditional and conditional predictions:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{v}_t, \boldsymbol{t}) \coloneqq w\boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t, \boldsymbol{t}) + (1-w)\boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t), \quad (12)$$

where w is the guidance weight to control the effect of conditional guidance, and $\tilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{v}_t, \boldsymbol{t})$ is used in Equation (11) for sampling generations with classifier-free guidance.

Two generators are used in our framework, *i.e.*, image-to-text embedding generator and text-to-image embedding generator. Equation (11) and Equation (12) describe the process of generating image embeddings given a text embedding as guidance. A similar process of generating text embeddings with an image embedding condition is performed by the text-to-image embedding generator in our framework.

Multiple Samplings. We fix the pretrained generators and generate embedding samples online to guide the representation learning process, which is regarded as embedding-level data augmentation. Given an input image-text pair for representation learning, we sample K embedding samples using the corresponding generator by translating the embeddings of input image and text. Multiple samplings readopted in the proposed framework (*i.e.*, $K > 1$). For example, given a text embedding $\boldsymbol{t}$ as a condition for text-to-image generator, we sample $\boldsymbol{v}'_1, \boldsymbol{v}'_2, \ldots, \boldsymbol{v}'_K$ generated embeddings online to guide the training of encoders.

Multiple samplings are beneficial to representation learning. First, sampling from diffusion generative models is a random process, which may produce hard samples. Multiple samplings could make the representation learning stable with different types of generated samples. Second, generating multiple samples from the generative distribution in the training batch provides more effective augmentation data in a single training step, which could accelerate the representation learning process.

The speed of sampling should be fast since we generate the embeddings online during training. DDIM (Song et al., 2021) sampling strategy is used to speed up the generation process. In practice, we observe that a small sampling step number (*e.g.* 5-10 steps) is sufficient to generate a valid embedding with favorable image-text alignment.

3.3. Generative Distribution Guidance

The generated embeddings from a diffusion-based generator are served as augmented embedding-level samples, which lie in a generative distribution. The number of generated samples could be infinite and be used to extend the limited real-world training data.

Given input image embedding $\boldsymbol{v}$ and text embedding $\boldsymbol{t}$, we generate image embeddings $\boldsymbol{v}'_1, \boldsymbol{v}'_2, \ldots, \boldsymbol{v}'_K$ by a text-to-image embedding generator conditioned on text embedding $\boldsymbol{t}$. Similarly, generated text embeddings $\boldsymbol{t}'_1, \boldsymbol{t}'_2, \ldots, \boldsymbol{t}'_K$ are sampled by an image-to-text embedding generator. The generated image embedding $\boldsymbol{v}'_k$ is viewed as a translation of the input text embedding $\boldsymbol{t}$. Aligning the input image embedding $\boldsymbol{v}$ and the generated image embeddings would implicitly align $\boldsymbol{v}$ with $\boldsymbol{t}$ from the input image-text pair. Therefore, we adopt contrastive loss to align the input image embedding $\boldsymbol{v}_i$ with corresponding generated image embeddings $\{\boldsymbol{v}'_{r1}, \ldots, \boldsymbol{v}'_{rk}, \ldots, \boldsymbol{v}'_{rK}\}$ from all positive text embeddings $\{\boldsymbol{t}_r\}_{r \in \mathcal{R}(i)}$:

$$\mathcal{L}_{i2i} = -\sum_{i=1}^{B}\sum_{k=1}^{K} \frac{1}{|\mathcal{R}(i)|} \sum_{r \in \mathcal{R}(i)} \log \frac{\exp(\boldsymbol{v}_i^T \boldsymbol{v}'_{rk}/\tau)}{\sum_{j=1}^{N} \exp(\boldsymbol{v}_i^T \boldsymbol{v}'_{jk}/\tau)}. \quad (13)$$

Similarly, the alignment from input text embedding to the set of generated text embeddings is defined as:

$$\mathcal{L}_{t2t} = -\sum_{j=1}^{B}\sum_{k=1}^{K} \frac{1}{|\mathcal{R}(j)|} \sum_{r \in \mathcal{R}(j)} \log \frac{\exp(\boldsymbol{t}_j^T \boldsymbol{t}'_{rk}/\tau)}{\sum_{i=1}^{N} \exp(\boldsymbol{t}_j^T \boldsymbol{t}'_{ik}/\tau)}. \quad (14)$$

Given the input embeddings and generated embeddings, the final learning objective is defined as:

$$\mathcal{L} = (\mathcal{L}_{i2t} + \mathcal{L}_{t2i}) + \lambda(\mathcal{L}_{i2i} + \mathcal{L}_{t2t}), \quad (15)$$

where λ is the weight to balance the losses.

The representation learning objective is composed of two main parts: embedding alignment from real-world data and generative distribution-guided learning. Using the generator as guidance can be also considered as training vision-language encoders by distilling knowledge from the generative distribution space.

4. Implementation Details

4.1. Model architecture.

The proposed framework is composed of four parts, an image encoder, a text encoder, an image-to-text embedding generator, and a text-to-image embedding generator. Following (Radford et al., 2021), we employ a widely-used vision Transformer ViT-B/32 (Dosovitskiy et al., 2021) as the image encoder, and the text encoder is a BERT-like 12-layer Transformer (Devlin et al., 2019) with width 512. We

Table 1. Performance with and with/o generator guidance and multiple-sampling (Multi-Sampling) on ImageNet dataset.

Generator	Multi-Sampling	Top-1 Acc.	Top-5 Acc.
x	x	30.1%	53.5%
✓	x	36.7%	61.3%
✓	✓	**39.1%**	**63.8%**

Table 2. Performance with and with/o generator guidance and multiple-sampling (Multi-Sampling) on COCO (C.) and Flick30K (F.) datasets for text-to-image (T2I) and image-to-text (I2T) retrieval tasks.

Generator	Multi-S.	C. T2I	C. I2T	F. T2I	F. I2T
x	x	11.0%	17.3%	17.8%	30.5%
✓	x	13.3%	21.0%	22.5%	37.9%
✓	✓	**14.5%**	**23.5%**	**26.2%**	**41.2%**

apply a projector to obtain the image or text embedding upon the output of the encoder, which is a two-layer MLP but is ignored in Section 3.2 for simplicity.

We adopt the diffusion prior model in DALL-E 2 (Ramesh et al., 2022) as the embedding generator. The generator is a 12-layer decoder-only Transformer with the input of image/text embeddings and predicts the text/image embeddings. We use the publicly available reproduction repository (LAION-AI, 2022) of pre-training DALL-E 2 model from LAION (Schuhmann et al., 2021). The image2text generator and text2image generator are pretrained on LAION-400M (Schuhmann et al., 2021) dataset with the embeddings predicted by a publicly available pretrained CLIP model (Radford et al., 2021). We note that the pretrained generators can be prepared in advance before training the proposed RLEG model.

For the process of input image-text pair, the image is resized to 224×224 and results in 7×7 image patches when the patch size is 32×32 for ViT-B/32. The text is truncated to 77 tokens as a pre-process. The number of multiple samplings K is set to 4 as a trade-off of speed and performance. The condition weight w during sampling is set to 2.0 for better image-text alignment, and dynamic thresholding is used following (Saharia et al., 2022). The loss weight λ is empirically set to 0.1.

4.2. Training settings.

We train the proposed model on the dataset of YFCC-15M used in CLIP (Radford et al., 2021), a subset of YFCC-100M (Thomee et al., 2016). Following the settings of (Radford et al., 2021), we use AdamW (Loshchilov & Hutter, 2019) as an optimizer, and the learning rate is initially set to $5e-4$ and decayed to zero with a cosine scheduler. A warm-up of the learning rate is used at the first 3 epochs. The weight decay for model parameters is 0.1. The model is trained from scratch for 32 epochs on 8 NVIDIA A100 GPUs. The batch size is set to 512 for each GPU card and a total of 4096 in the experiments.

4.3. Downstream tasks.

After pre-training, the proposed model is evaluated in a zero-shot setting on several downstream datasets, including ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), and Flickr30K (Young et al., 2014). We note that the generators in our framework are only used during pre-training, and the embeddings extracted from the image encoder and text encoder are finally adopted for downstream tasks.

For the image classification task of ImageNet, we construct prompts with the $1K$ class label names following the setting in CLIP (Radford et al., 2021). Then $1K$ text embeddings are extracted from these text inputs with class label names. Given an input image, the distances from image embedding to the text embeddings are computed, and the class label is predicted by the closest distance. Top-1 accuracy and top-5 accuracy are used for evaluation.

For the zero-shot image-text retrieval on MS-COCO and Flickr30K, the pretrained model is used to extract embeddings from images and texts separately. Similarity scores between image embeddings and text embeddings are used for ranking. We use the R@K to report the recall of top-K retrieval items.

We also conduct experiments on other downstream tasks including object detection, semantic segmentation, and text-conditional image generation, which will be described in each experiment section.

5. Experiments and Analysis

We first conduct several ablation studies to show the comparisons when using different settings in the proposed framework and give analysis on the help of introducing generative models into representation learning . The proposed method is evaluated on different downstream tasks, including image classification, image-text retrieval, object detection, semantic segmentation, and text-conditional image generation.

All the models in the experiments are trained on YFCC-15M (CLIP version (Radford et al., 2021)) for a fair comparison. Comparisons with previous methods are then provided on image classification and image-text retrieval tasks.

5.1. Ablation Study

Generative models help vision-language learning. We evaluate the models on both the image classification task on ImageNet-1K (Deng et al., 2009) and the image-text retrieval task on MS-COCO (Lin et al., 2014) and Flicker30K (Young et al., 2014) datasets.

From Table 1, we observe that with the help of the guid-

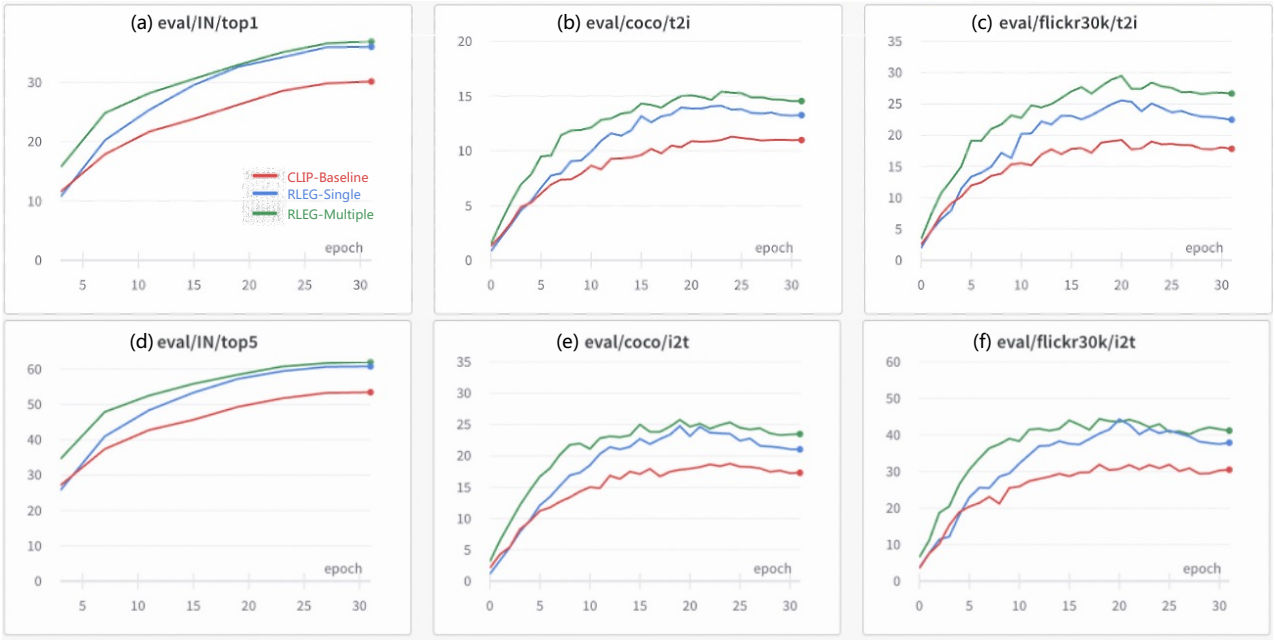

Figure 3. Comparisons during the training process among the methods without generator guidance (CLIP-Baseline), single sampling guidance (RLEG-Single), and multiple samplings guidance (RLEG-Multiple). Thanks to generative embeddings, the training process is accelerated to achieve similar results with smaller training epochs.

ance of diffusion-based generators, the proposed method outperforms the baseline with respectively $+9.0\%$, $+10.3\%$ improvements on top-1 and top-5 accuracies of ImageNet-1K classification in the zero-shot setting. The generated embeddings provide strong and semantic data augmentation for each class concept, which would boost the model to learn a more robust representation for image classification.

For the image-text retrieval task, significant improvements on COCO and Flicker30K are achieved when learning with embedding generation, as shown in Table 2. Specifically, the proposed method surpasses the baseline with $+3.5\%$, $+6.2\%$, $+8.4\%$, $+10.7\%$ on text-to-image and image-to-text retrieval top-1 accuracy of COCO and Flicker30K, respectively. The embedding generators act as modality translators during the sampling process. Training with generative distribution guidance could implicitly force the model to align the image and text embeddings, which is beneficial to the cross-modal downstream tasks.

Furthermore, the representation learning process could be accelerated with the help of sampling more effective data from the generative distribution in one training step. As shown in Figure 3, the model with generation guidance can achieve similar results with smaller training epochs compared to the baseline method (*i.e.*, CLIP).

Generative guided model is a better generator backbone. The vision-language learning with generator guidance could produce a more powerful model with rich semantic information, which is advantageous to train a better image generator.

Table 3. Performances with and with/o generator guidance on ImageNet 64×64 and COCO 64×64 image generation tasks.

Method	Generator	ImageNet FID	COCO FID
CLIP Backbone	x	11.5	15.7
RLEG Backbone	✓	**9.3**	**13.1**

Table 4. Performances with and with/o generator guidance on COCO object detection and segmentation tasks in zero-shot setting.

Method	Generator	Detection (AP)	Segmentation (mIoU)
CLIP	x	42.7	7.9
RLEG	✓	**46.9**	**11.3**

We follow the DALL-E 2 framework and experimental settings to train image generators with different representation backbones of baseline CLIP and the proposed model.

We evaluate the performance of image generation on dataset ImageNet and COCO in 64×64 generated image resolution. The results of FID metric (Heusel et al., 2017) are shown in Table 3. We can observe that the proposed method equipped with embedding generators during training achieves better results than the CLIP-based image generator. Learnt generative model enforces the image encoder to focus on more semantic information using a large augmented embedding space as a distillation target, resulting in a better representation backbone for training the generator.

Generative augmentation helps dense tasks. The generated samples from a dense generative distribution could

Table 5. Comparison with the combination of different dataset size and sampling factor K. ImageNet Top-1 accuracy (%) is reported, and the percentage numbers in parentheses are the relative improvements based on K=0.

Dataset	K=0	K=1	K=2	K=4
YFCC3M	16.4	20.1 (+22.6%)	22.7 (+37.6%)	**24.3** (+48.2%)
CC3M	18.1	21.7 (+19.9%)	23.9 (+32.0%)	**25.8** (+42.5%)
CC12M	26.8	31.5 (+17.5%)	34.1 (+27.2%)	**35.6** (+32.3%)
YFCC15M	30.1	36.7 (+21.9%)	37.8 (+25.6%)	**39.1** (+29.9%)

provide more fine-grained information for learning a better image encoder, which is beneficial to apply on dense downstream tasks, such as object detection and semantic segmentation.

Following the setting in RegionCLIP (Zhong et al., 2022), zero-shot object detection is conducted by matching the features within proposal boxes (ground-truth boxes are used for simplicity) and the class labels. It is used to evaluate the representation learnt on local patches when applied to object detection tasks. The results shown in Table 4 demonstrate that the proposed model could attend to local patches and extract more effective representation, which is beneficial for object detection tasks.

To conduct zero-shot semantic segmentation, we apply a pixel-text matching following DenseCLIP (Zhou et al., 2022) by extracting embeddings from all segmentation class labels for local point classification on image feature map. To enlarge the resolution of the final image feature map, we use 336×336 image as input and divide it with 16×16 patch size for all compared models. From Table 4, we can see that the recognition ability on patch-level is improved with generator guidance for segmentation task. We conjecture that sampling from the generative distribution could provide more appearance similar objects and augment the samples with more fine-grained information.

Generating single or multiple samples. We use multiple samplings in the proposed framework to generate more embeddings for training. Experiments are conducted to compare the performance of multiple samplings and single sampling when generating image and text embedding samples. In a single sampling setting, only one embedding sample is generated (*i.e.*, $K = 1$) by each generator in our framework, and other settings are kept unchanged. The results in Table 1 and Table 2 show that sampling multiple embeddings in one training step could help learn more effective representation and accelerate the representation learning process as shown in Figure 3.

We also conduct several experiments with the combination of different dataset size and sampling factor K in Table 5. Specifically, we randomly select a subset of YFCC15M to form a 3M size dataset YFCC3M, and use CC3M and CC12M for another different size sources. For sampling factor K, we use four settings of K=0,1,2,4, where K=0 means

Table 6. Performance with large training dataset LAION-400M and large model size ViT-L/14 for image encode on ImageNet (IN), COCO (C.) and Flick30K (F.) datasets for image classificatin, text-to-image (T2I) and image-to-text (I2T) retrieval tasks.

Method	IN Top 1	C. T2I	C. I2T	F. T2I	F. I2T
CLIP	75.3	36.1	57.3	64.6	85.7
RLEG	**79.8**	**43.7**	**59.5**	**73.9**	**89.3**

Table 7. Comparison with different training time of CLIP and RLEG on ImageNet-1K image classification task, where 16E and 32E mean training with 16 epochs or 32 epochs.

Method	CLIP-16E	RLEG-16E	CLIP-32E	RLEG-32E
ImageNet Top-1	24.8%	33.2%	30.1%	**39.1%**
Training Time	9.6h	17.6h	19.2h	35.2h

the traditional CLIP model without generator guidance. In addition, more experiments with larger K are not conducted because of the limited GPU memory. As shown in Table 5, sampling more samples consistently obtain superior performance on different size datasets. The relative improvements of K=1 are similar for YFCC3M and YFCC15M, while significant improvements are observed on YFCC3M and CC3M when K=4. It demonstrates that smaller pre-training datasets benefit more from more augmentations, which is reasonable since data lacking problem plays more important role for smaller dataset and enriching the training data could improve the performance significantly.

Training with larger dataset and model. To evaluate the scalability of the proposed framework, we train the proposed model on a larger dataset LAION-400M (Schuhmann et al., 2021) with a large ViT-L/14 encoder following the settings in CLIP (Radford et al., 2021). The results in Table 6 show that the proposed method still outperforms CLIP baseline on ImageNet classification and image-text retrieval tasks when pretraining on large-scale dataset. The improvements demonstrate that training with generative distribution guidance is scalable to be effective for representation learning.

Different hyper-parameter settings. We study different settings of hyper-parameters used in our framework on ImageNet top-1 accuracy. First, when using identity mapping as the projector for mapping the embedding instead of a two-layer MLP, the performance drops from 0.7% to 38.4%. Second, sampling embeddings with 5, 10, 50 steps in DDIM strategy obtains slightly different results of $38.7\%, 39.1\%, 39.2\%$. Third, the condition weight w during sampling with $0.1, 1.0, 2.0, 5.0$ attains $32.4\%, 38.3\%, 39.1\%, 38.8\%$ results, which controls the balance of diversity and alignment. We note that the model always promotes the performance compared to the baseline without generator guidance under all different settings.

Number of parameters and training cost. The number of learnable parameters of RLEG and CLIP is the same during training. In testing phase, the generators are not used and

Table 8. Comparison with previous vision-language pretraining methods with different supervision and augmentation on both vision and vision-language tasks, including image classification on ImageNet-1K and image-text retrieval on COCO and Flickr30K. All the models are evaluated with the same backbone, dataset, and other training settings.

Method	ImageNet		COCO T2I		COCO I2T		Flickr30K T2I		Flickr30K I2T	
	Top-1	Top-5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP (Radford et al., 2021)	30.1	53.5	11.0	26.3	17.3	34.4	17.8	30.0	30.5	44.2
SLIP (Mu et al., 2022)	32.5	55.9	12.3	28.1	18.9	37.6	19.5	33.9	32.5	47.9
MS-CLIP (You et al., 2022)	34.3	57.2	13.1	29.9	20.3	43.2	22.4	40.1	34.8	51.3
MaskCLIP (Dong et al., 2023)	36.0	58.8	14.1	34.1	21.8	44.6	24.2	43.5	38.9	57.0
DeCLIP (Li et al., 2022b)	36.2	59.3	14.3	34.5	21.6	44.1	24.6	45.8	39.2	58.7
RLEG (Ours)	**39.1**	**63.8**	**14.5**	**34.7**	**23.5**	**47.7**	**26.2**	**44.2**	**41.2**	**60.1**

the inference time is also same for RLEG and CLIP.

In the experiments, training one epoch for RLEG and CLIP costs 1.1h and 0.6h respectively. To compare the performances of RLEG and CLIP with same training time, we train RLEG with about 16 epochs and CLIP with 32 epochs. As shown in Table 7, RLEG still outperforms CLIP on ImageNet Top-1 accuracy (33.2% vs. 30.1%) with similar training cost. For a fair comparison, we also report the result of CLIP when training with 16 epochs, which achieves 24.8% as shown in Table 7.

5.2. Comparison with Previous Methods

In this section, we compare RLEG with previous vision-language pretraining methods, including the baseline CLIP (Radford et al., 2021), the variants of SLIP (Mu et al., 2022), MS-CLIP (You et al., 2022), DeCLIP (Li et al., 2022b) and MaskCLIP (Dong et al., 2023). Since the reported results of these methods are obtained with different backbones and dataset settings, we re-implement these methods by using the same settings as ours for a fair comparison. Experiments are conducted on both vision and vision-language tasks, including image classification on ImageNet (Deng et al., 2009) and image-text retrieval on COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014).

Zero-shot image classification. The ImageNet classification results in Table 8 show that our method outperforms state-of-the-art vision-language pretraining methods with a new learning scheme. The methods of SLIP, DeCLIP, and MaskCLIP utilize more supervision signals from the available input dataset, such as self-supervision by multi-view augmentation or patch-masked augmentation. In contrast, the proposed method performs an embedding-level data augmentation by generating samples instead of modifying the input image or text, which extends the semantic content from a dense generative distribution beyond the training dataset. The augmented data could fulfill the concepts of classes, which is beneficial to the representation learning for image classification.

Zero-shot image-text retrieval. We further compare the methods for zero-shot image-text retrieval task on MS-COCO (Lin et al., 2014) and Flicr30K (Young et al., 2014). The image-to-text and text-to-image results are reported in Table 8. We report R@1 and R@5 metrics for both two settings. We can observe from the results that the proposed method achieves a favorable performance for the cross-modal retrieval task. The embedding generators used in our framework serve as translators between image and text modalities, which implicitly improves the image-text alignment during the representation learning process.

6. Limitations and Social Impacts

Although RLEG improves the representation learning on many downstream tasks, it is limited to do alignment tasks on image and text inputs, which is not as flexible as generative methods like image captioning. The diversity of augmentation is also limited to the capability of the pretrained diffusion generators. Before used in production, further analysis of the data and model should be taken to reduce social biases. For example, harmful input texts or unsuitable images may affect the model to produce unwanted results in the real world.

7. Conclusion

We have presented a simple but effective representation learning method, named RLEG, guided by diffusion-based embedding generators. The diffusion models are exploited to generate embeddings online to help learn effective vision-language representation. The pretrained generators transfer the embedding modality online between vision and language domains. The generated embeddings are served as augmented embedding-level samples, which are then applied to contrastive learning. The method could also be considered as training vision-language encoders by distilling knowledge from the generative data space. Experimental results validate the effectiveness of the proposed RLEG on various tasks including image classification, cross-modal retrieval, object detection, semantic segmentation and text-conditional image generation.

References

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R. N., Hammers, A., Dickie, D. A., del C. Valdés Hernández, M., Wardlaw, J. M., and Rueckert, D. GAN augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. COYO-700M: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Cai, R., Yang, G., Averbuch-Elor, H., Hao, Z., Belongie, S., Snavely, N., and Hariharan, B. Learning gradient fields for shape generation. In *ECCV*, pp. 364–381, 2020.

Dai, Z., Chen, M., Gu, X., Zhu, S., and Tan, P. Batch dropblock network for person re-identification and beyond. In *ICCV*, pp. 3691–3701, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. In *NeurIPS*, pp. 8780–8794, 2021.

Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., Wen, F., and Yu, N. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, pp. 10995–11005, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pp. 6629–6640, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(47):1–33, 2022a.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. In *NeurIPS*, pp. 8633–8646, 2022b.

Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., and Dittadi, A. Diffusion models for video prediction and infilling. *TMLR*, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916, 2021.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of StyleGAN. In *CVPR*, pp. 8110–8119, 2020.

Kumar, V., Glaude, H., de Lichy, C., and Campbell, W. A closer look at feature space data augmentation for few-shot intent classification. In *DeepLo Workshop*, pp. 1–10, 2019.

LAION-AI. Pretrained DALL-E 2 from LAION. https://github.com/LAION-AI/dalle2-laion, 2022.

Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., and DUAN, L. Uncertainty modeling for out-of-distribution generalization. In *ICLR*, 2022a.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Liu, W., Ma, L., and Cui, M. Learning-based stereoscopic view synthesis with cascaded deep neural networks. *JACIII*, 26(3):393–406, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Mehrjou, A., Schölkopf, B., and Saremi, S. Annealed generative adversarial networks. *arXiv preprint arXiv:1705.07505*, 2017.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. SLIP: Self-supervision meets language-image pre-training. In *ECCV*, pp. 529–544, 2022.

Mumuni, A. and Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *ICML*, pp. 8162–8171, 2021.

Niu, C., Song, Y., Song, J., Zhao, S., Grover, A., and Ermon, S. Permutation invariant graph generation via score-based generative modeling. In *ICAIS*, pp. 4474–4484, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Gontijo-Lopes, R., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pp. 36479–36494, 2022.

Sajjadi, M. S., Parascandolo, G., Mehrjou, A., and Schölkopf, B. Tempered adversarial networks. In *ICML*, pp. 4451–4459, 2018.

Schuhmann, C., Kaczmarczyk, R., Komatsuzaki, A., Katta, A., Vencu, R., Beaumont, R., Jitsev, J., Coombes, T., and Mullis, C. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, 2021.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. FLAVA: A foundational language and vision alignment model. In *CVPR*, pp. 15617–15629, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *ICLR*, 2021.

Song, Y. and Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. In *NeurIPS*, pp. 11918–11930, 2020.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. Neural discrete representation learning. In *NeurIPS*, pp. 6306–6315, 2017.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. GIT: A generative image-to-text transformer for vision and language. *TMLR*, 2022a.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022b.

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. Unified contrastive learning in image-text-label space. In *CVPR*, pp. 19163–19173, 2022.

You, H., Zhou, L., Xiao, B., Codella, N., Cheng, Y., Xu, R., Chang, S.-F., and Yuan, L. Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, pp. 69–87, 2022.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. CoCa: Contrastive captioners are image-text foundation models. *TMLR*, 2022.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zhang, L., Jiang, N., Diao, Q., Zhou, Z., and Wu, W. Person re-identification with pose variation aware data augmentation. *Neural Comput. Appl.*, 34(14):11817–11830, 2022.

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., and Gao, J. RegionCLIP: Region-based language-image pretraining. In *CVPR*, pp. 16793–16803, 2022.

Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from CLIP. In *ECCV*, pp. 696–712, 2022.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pp. 13041–13049, 2020.