

DiffDis: Empowering Generative Diffusion Model with Cross-Modal Discrimination Capability

Runhui Huang¹ Jianhua Han² Guansong Lu² Xiaodan Liang^{1*}
Yihan Zeng² Wei Zhang² Hang Xu²

¹Shenzhen Campus of Sun Yat-sen University ²Huawei Noah’s Ark Lab

Abstract

Recently, large-scale diffusion models, e.g., Stable diffusion and Dalle2, have shown remarkable results on image synthesis. On the other hand, large-scale cross-modal pre-trained models (e.g., CLIP, ALIGN, and FILIP) are competent for various downstream tasks by learning to align vision and language embeddings. In this paper, we explore the possibility of jointly modeling generation and discrimination. Specifically, we propose **DiffDis** to unify the cross-modal generative and discriminative pretraining into one single framework under the diffusion process. DiffDis first formulates the image-text discriminative problem as a generative diffusion process of the text embedding from the text encoder conditioned on the image. Then, we propose a novel dual-stream network architecture, which fuses the noisy text embedding with the knowledge of latent images from different scales for image-text discriminative learning. Moreover, the generative and discriminative tasks can efficiently share the image-branch network structure in the multi-modality model. Benefiting from diffusion-based unified training, DiffDis achieves both better generation ability and cross-modal semantic alignment in one architecture. Experimental results show that DiffDis outperforms single-task models on both the image generation and the image-text discriminative tasks, e.g., 1.65% improvement on average accuracy of zero-shot classification over 12 datasets and 2.42 improvement on FID of zero-shot image synthesis.

1. Introduction

“What I cannot create, I do not understand.” by Richard Feynman (a well-known theoretical physicist).

Recently, large-scale diffusion models (DM) [18, 48, 55] such as Stable Diffusion [45] and Dalle2 [43] have shown impressive results in image synthesis and re-define the capacity of state-of-the-art text-guided image synthesis. Typ-

*Corresponding author: xdliang328@gmail.com

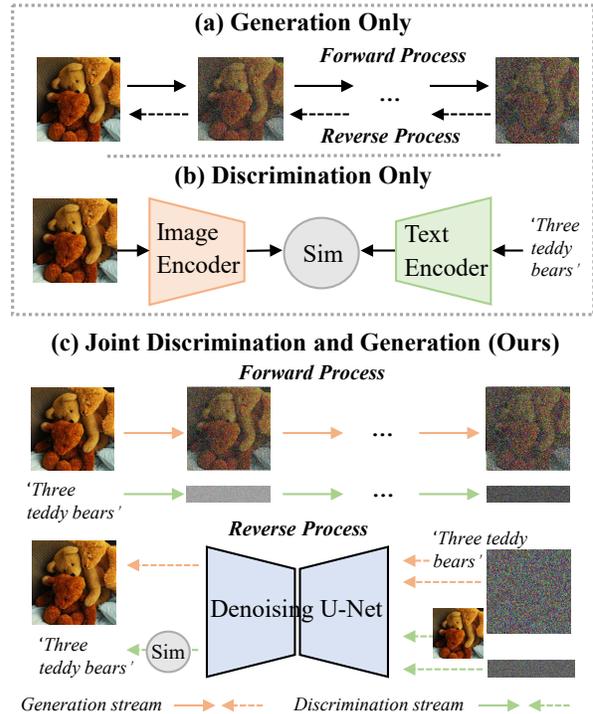


Figure 1. Comparison of our framework and single-task models. (a) The diffusion-based image generation-only model. (b) The image-text discrimination-only model. (c) Our DiffDis joints the discriminative and generative tasks under the diffusion processing into one framework. Better viewed in colors.

ically speaking, these models contain more than one billion parameters and have a large model capacity; thus have a good generalization and cover an extensive range of domains. Here, by rethinking the famous remarks from Richard Feynman, we explore *whether such powerful generative models can learn the ability to further discriminate and understand cross-modal data*.

On the other hand, recent large-scale Vision-Language Pre-training (VLP) models [42, 22, 57, 28] like CLIP[42] and ALIGN [22] have demonstrated success in vari-

ous downstream zero-shot image classification or retrieval tasks. Similar to large-scale diffusion models, these models are pre-trained with millions of image-text pairs collected from the Internet. The critical idea of these works is to contrastively align the image and text embeddings into a joint feature space, thus gaining zero-shot discrimination capability, which is different from the diffusion models that consider the problem as the parameterized Markov chain. In this paper, we focus on bridging the generative diffusion models with VLP models to empower the generative diffusion model with the cross-modal discrimination capability, in the spirit of similar principles via large-scale pretraining.

There exist some methods that have considered combining generative models and discriminative models into a single framework. The famous generative adversarial networks (GAN) [15] introduce the discriminator to guide the adversarial learning of the generator. However, the implicit adversarial of GAN would lead to mode-collapse and training instabilities. Moreover, HybViT [56] tried to replace the UNet structure in GLIDE [36] with a ViT [14] model and directly added a classification head to perform image generation and classification jointly, while ignoring the powerful diffusion process for the discriminate task. On the other hand, recent unified general vision models such as Pix2Seq [6], OFA [54], and Unified I/O [34] try to unify different vision-language tasks into an autoregressive sequence prediction framework. However, the autoregressive image generation solutions such as Dalle [44] and OFA [54] show inferior performance compared to the DM-based models, such as Dalle2 [43] and Stable Diffusion [45], in terms of both generation quality and sampling efficiency.

In this paper, we present **DiffDis**, a unified vision-language diffusion model for both generative and discriminative tasks under the diffusion paradigm. Specifically, DiffDis first formulates the image-text discriminative problem as a generative diffusion process of the text embedding outputted by the text encoder conditioned on the input image. Therefore, the generation and discrimination tasks can share the same image-branch network (i.e., original U-Net) in the multi-modality diffusion model. During inference, zero-shot image classification is performed by calculating the cosine similarity between the generated text embedding and the downstream text embeddings. Secondly, we design a dual-stream network architecture to better fuse the knowledge of latent images with different scales into the text query in image-text alignment. Finally, a unified training paradigm is further proposed to alternatively feed the required inputs and the conditions when jointly performing diffusion-based generative and discriminative tasks. When training discriminative tasks, the image branch serves as an image encoder to feed conditional information into the reverse text embedding diffusion process and vice versa.

Extensive experiments have shown that the proposed

DiffDis method can achieve better performance on both zero-shot classification and text-guided image generation tasks. Compared to the single-task baseline, our unified framework DiffDis can achieve 1.65% improvement on the average zero-shot classification accuracy on 12 datasets and a 2.42 improvement on FID of zero-shot image synthesis compared to the single-task model. This work is the first to unify the training of generative and discriminative tasks under the diffusion process. We hope that this research will serve as an early-stage exploration for future studies aiming to unify these two tasks under the diffusion process, thereby providing more choices for future multi-task multi-modal jointly-training frameworks.

Our contributions can be summarized as:

- We propose DiffDis to explore a unified vision-language diffusion model for both multi-modality generation and discrimination tasks.
- DiffDis reformulates the image-text discriminative problem by utilizing a generative diffusion process of the text embeddings conditioned on input images.
- We propose a dual-stream network architecture and a diffusion-based unified training paradigm for jointly training the generative and discriminative tasks.
- Extensive experiments demonstrate that our DiffDis outperforms single-task models, achieving a 1.65% improvement on average zero-shot classification accuracy across 12 datasets and a 2.42 improvement on FID of text-guided image generation. Additionally, DiffDis outperforms CLIP, with a 4.7% improvement on average zero-shot classification accuracy across 12 datasets and a 14.5% improvement on average R@1 of image-text retrieval tasks on Flickr30k and MSCOCO.

2. Related Work

Vision-Language Pre-training. Current communities in natural language processing and computer vision both favor the pre-train-and-fine-tune scheme because of the superior performance of the pre-trained models [11, 4, 13]. Recent works such as CLIP[42] and ALIGN [22] then extend this diagram to a joint cross-modal domain of Vision-and-Language Pre-training (VLP). These large-scale models have shown promising results in various downstream tasks, such as zero-shot classification and image-text retrieval. Their great generalization ability mainly comes from the large-scale automatic-collected image-text dataset from the Internet (e.g, YFCC100M [51], CC12M [5]) The VLP models can be categorized by their pre-training tasks: (a) Image-text contrastive learning task: CLIP [42], ALIGN [22], FILIP [57] and UNIMO [28] utilize the cross-modal

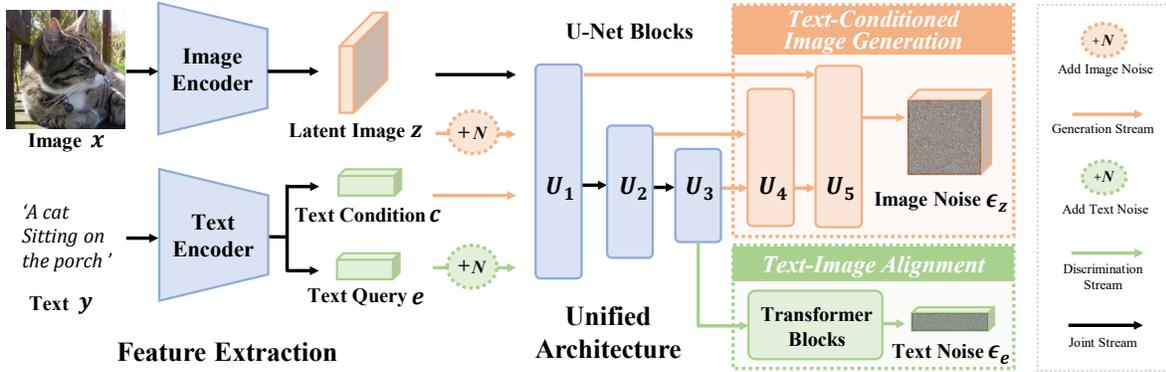


Figure 2. Overall model architecture of DiffDis. DiffDis includes an image encoder to encode the image input x into the latent image z and a text encoder to obtain the text condition c and text query e with the caption input y . For text-conditional image generation, the latent image z will be added noise and is fed into the UNet with c as the condition to predict the added noise ϵ_z . For image-text alignment learning, the text query e will be added noise and is fed into the UNet with latent image z as the condition to predict the added noise ϵ_e .

contrastive learning which aligns the textual and visual embedding; (b) Language Modeling (LM) based tasks: VisualBERT [27], UNITER [7], M6[29], and DALL-E [44] employ LM-like objectives, including both masked LM (e.g., Masked Language), and autoregressive LM (e.g., image captioning). In contrast, we try to follow the diffusion framework to perform image-text alignment learning in an end-to-end unified manner while maintaining the benefit of strong image generation ability.

Denosing Diffusion Probabilistic Models. Since Ho *et al.* [18] build a connection between diffusion model [48] and denosing score matching model [50] and propose DDPMs (Denosing Diffusion Probabilistic Models) to achieve high image generation quality, diffusion models start to attract attention. Nichol *et al.* [37] propose to learn the variances of the reverse diffusion process to achieve higher sampling efficiency with fewer forward passes. Dhariwal *et al.* [12] achieve better image generation quality than GANs by finding a better model architecture through ablations and propose a new sampling technique called classifier guidance. While classifier guidance requires training an extra classifier model, Ho *et al.* [20] propose classifier-free guidance to circumvent this problem by jointly training a conditional and an unconditional model and combine the resulting conditional and unconditional scores to achieve the same effect as classifier guidance. Recently, diffusion models are applied to text-to-image generation and achieve appealing generation results [36, 19, 43, 46, 45]. Some methods [56, 1] make an attempt to unify the generation and discriminative tasks by directly adding a classification head into the UNet structure while our DiffDis formulate the discriminative problem into the powerful diffusion process.

3. Methodology

In this section, we first review some preliminaries about diffusion models (Sec. 3.1). Then we introduce our formu-

lations of generative and discriminate tasks under the unified diffusion process (Sec. 3.2), followed by a detailed description of the network architecture of the proposed unified DiffDis (Sec. 3.3). Finally, the training paradigm is introduced to deal with the generative and discriminative tasks with different inputs and conditions (Sec. 3.4).

3.1. Preliminary on Diffusion Models

Given a sample from the real data distribution $x_0 \sim q(x_0)$, Gaussian diffusion models first produce a Markov chain of latent variables x_1, \dots, x_T by progressively adding Gaussian noise to the sample according to some variance schedule given by β_t as follows:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \quad (1)$$

and then learn a model to approximate the true posterior:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (2)$$

to perform the reverse denosing process for sampling: starting from a random noise $x_T \sim \mathcal{N}(0, I)$ and gradually reducing the noise to finally get a sample x_0 . While a tractable variational lower-bound \mathcal{L}_{VLB} on $\log p_\theta(x_0)$ can be used to optimize μ_θ and Σ_θ , to achieve better results, Ho *et al.* [18] instead adopt a denosing network $\epsilon_\theta(x_t, t)$ which predicts the noise component of a noisy sample $x_t \sim q(x_t | x_0)$ and the following training objective:

$$\mathcal{L} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0, I), t \sim [1, T]} \|\epsilon - \epsilon_\theta(x_t, t)\|^2, \quad (3)$$

where t uniformly sampled from $\{1, \dots, T\}$.

3.2. Task Reformulation

Basic Notations. We denotes the image-text dataset as $D = \{x_i, y_i\}_{i=1}^N$, where x_i and y_i denote the i -th image and text.

N is the total number of image-text pairs. As shown in Fig. 2, DiffDis consists of an image encoder \mathcal{V} and text encoder \mathcal{T} . Besides, Φ_u and Φ_θ represent the UNet model used for text-conditional image generation, and the encoder part of UNet model with additional transformer blocks for image-text alignment learning, respectively.

Diffusion-based Text-conditioned Image Generation.

The generative part of DiffDis aims to generate images conditioned on input text prompts. Following the standard diffusion models [36, 19, 43, 46, 45], DiffDis generates image samples by gradually removing the noise from a random Gaussian noise signal over a finite number of steps. The diffusion model is trained by adding and predicting the different levels of noise on the image in the opposite direction of the sampling process.

Specifically, in the training procedure, the diffusion process of image can be represented as a parameterized Markov chain, which adds T steps' random Gaussian noise ϵ to gradually convert the original image x_0 to a random Gaussian distribution x_T . Following the LDM [45], we utilize the latent image $z = \mathcal{V}(x) \in \mathbb{R}^{H \times W \times d_z}$ outputted by the image encoder \mathcal{V} instead of the RGB space of image x as the input signal. Therefore, the diffusion-based text-conditional image generation loss \mathcal{L}_{IG} , which aims to predict the added Gaussian noise, can be formulated as:

$$\mathcal{L}_{IG} = \mathbb{E}_{\mathcal{V}(x), \epsilon_z \sim \mathcal{N}(0, I), t_z} [\|\epsilon_z - \Phi_u(z_t, t_z, c)\|^2] \quad (4)$$

where text condition $c = \mathcal{T}(y) \in \mathbb{R}^{L \times d_y}$ denotes the token-wise representations of the text prompt y . L and d_y represent the context length and embedding dimensions of outputted token embeddings, respectively.

During the inference (sampling) process, starting from a random Gaussian noise $z_T \sim \mathcal{N}(0, I)$, we reverse the diffusion process and gradually remove the predicted noise $\Phi_u(z_t, t_z, c)$ to obtain the sampled latent image after finite steps. We use an image decoder \mathcal{D} to convert the sampled latent image back to RGB space $\hat{x} = \mathcal{D}(\hat{z}_0)$. DDIM sampler [49] and classifier-free guidance [20] are employed. More details can refer to Algorithm 2.

Diffusion-based Image-text Alignment Pretraining. Previous methods perform image-text alignment pretraining by aligning the visual feature and textual feature in a common semantic space, where the positive image-text pairs are pulled towards each other and the negative image-text pairs are pushed against each other [57, 22]. The image-text contrastive (ITC) loss, like in CLIP [42], first calculates the cosine similarity of normalized visual feature v_i and textual feature e_j to measure the relevance of each image and text: $s_{i,j}^y = s_{i,j}^x = v_i^\top e_j$, where $s_{i,j}^x$, $s_{i,j}^y$ denote the image-to-text similarity and the text-to-image similarity. Besides, v_i is the global representation of the image x_i and e_j is the global representation of the text y_j . Based on $s_{i,j}^x$ and $s_{i,j}^y$, the ITC loss can be calculated as [42].

To reformulate this image-text alignment problem into a diffusion process, we propose the denoising diffusion-based image-text alignment which treats the latent image z as the condition and learns the distribution of the corresponding text embedding $e \in \mathbb{R}^{1 \times d_y}$. The diffusion process of e can be formulated as

$$e_t = \gamma \sqrt{\bar{\alpha}_t} e_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_e, \quad (5)$$

$$\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{j=0}^t \alpha_j, \quad (6)$$

where we omit the index i , and β_t is used to control the strength of added noise for timestep t_e . The $\gamma \in (0, 1]$ is the scale factor to scale the text embedding e_0 .

In contrast to the image generation task, the diffusion model Φ_θ is trained to estimate the original clean text query $\hat{e}_0 = \Phi_\theta(e_t, t_e, z)$. Note that Φ_θ can also be a noise prediction model to predict the noise ϵ_e .

Then the diffusion-based image-text alignment objective is to minimize the distance between \hat{e}_0 and e . In detail, we first calculate the cosine similarity between \hat{e}_0 and e :

$$s = \hat{e}_0^\top e, \quad (7)$$

Note that we omit the index of the embeddings here. Then the diffusion-based image-text alignment loss \mathcal{L}_{ITA} can be calculated as:

$$\mathcal{L}_{ITA} = -\frac{1}{2B} \sum_{i=1}^B \left[\log \frac{\exp(s_{i,i}^x)}{\sum_j \exp(s_{i,j}^x)} + \log \frac{\exp(s_{i,i}^y)}{\sum_j \exp(s_{j,i}^y)} \right], \quad (8)$$

where B denotes the batch size and \mathcal{L}_{ITA} is calculated for each diffusion step during training.

3.3. Unified Model Architecture

Feature Extraction Modules. As shown in Fig. 2, DiffDis includes an image encoder to obtain the latent image [45] and a text encoder to obtain the text condition and text query. Specifically, the image encoder \mathcal{V} of an autoencoder aims to convert the image x from RGB space to image's latent representation z which improves the training efficiency and perform image generation on high-resolution image synthesis. On the other hand, we adopt the text encoder \mathcal{T} to encode the text prompt y to the text condition c and text query e . Note that the dimensions of these two text embedding outputs are different. The text condition $c \in \mathbb{R}^{L \times d_y}$ is the token-wise representation of the text prompt y while the text query $e \in \mathbb{R}^{1 \times d_y}$ is the normalized global representation of the text prompt y .

Unified Architecture. The following UNet's structure receives three inputs: latent image z , text condition c and text query e . For *text-conditional image generation task*,

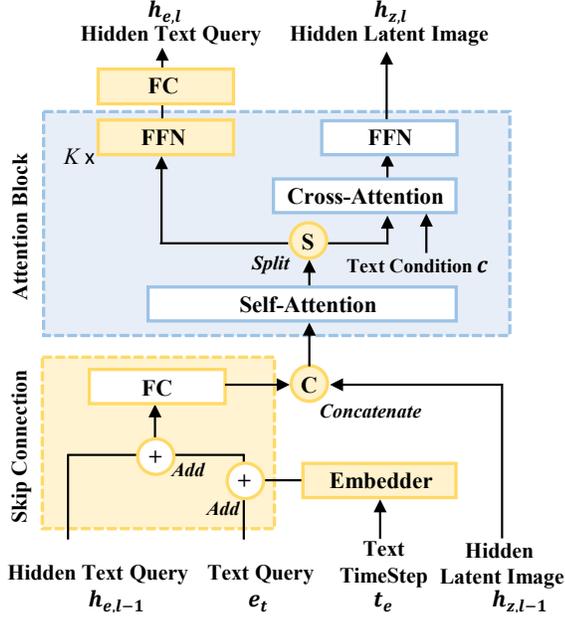


Figure 3. Detailed architecture of dual-stream deep fusion attention block. Compare to the conventional attention block (blue part) in [45], we design a dual-stream deep fusion attention architecture to better fuse the knowledge of latent image into the text query in cross-modal alignment learning. The separate FFNs learn modality-specific information. Besides, we build a cross-block skip connection from the noised text query e_t with time condition to the hidden text query $h_{e,l-1}$ outputted from the last block.

we adopt the noisy latent image z_t and text condition c as input to predict the noise ϵ_z added on latent image z . For *diffusion-based image-text alignment learning*, we adopt the latent image z and noised text query e_t as input to predict the original clean text query e . Note that it is viable to predict the noise ϵ_e added on text query e for the alignment learning. More training details can refer to Sec. 3.4. Besides, the Φ_θ also contains a unique transformer with M transformer block and a linear predictor. The unique transformer follows the middle block of the UNet to obtain more semantic information. The transformer’s input is the concatenation of the flattened image’s feature map and the text query outputted from the middle block of UNet. Finally, the linear predictor will be fed by the text query token and predict the original clean text query e .

Dual-stream Deep Fusion Attention Block. We modify the architecture of attention blocks in UNet to better unify these two tasks with different inputs. Previous stable diffusion [45] directly use K transformer decoder blocks to inject the text condition information into the latent image via the cross-attention mechanism. As illustrated in Fig. 3, we proposed a dual-stream deep fusion block to better fuse the latent image knowledge into the text query for cross-modal alignment learning. Specially, we adopt an additional fully-

Algorithm 1 Diffusion-base Unified Training

Input: Image x , Text y , Image Encoder \mathcal{V} , Text Encoder \mathcal{T} , Timestep T .

- 1: **repeat**
 - 2: $c, e = \mathcal{T}(y)$ ▷ Get text condition and text query
 - 3: $z = \mathcal{V}(x)$
 - 4: $t_z, t_e \sim \text{Uniform}(\{1, \dots, T\})$
 - 5: $\epsilon_z, \epsilon_e \sim \mathcal{N}(0, I)$
 - 6: Mask e and calculate \mathcal{L}_{IG} based on Eq. 4.
 - 7: Mask c and calculate \mathcal{L}_{ITA} based on Eq. 8.
 - 8: Calculate total loss \mathcal{L} based on Eq. 9
 - 9: Take gradient descent step on

$$\nabla_{u, \theta} \mathcal{L}_{Total}$$
 - 10: **until** converged
-

connected layer to project the text query embedding into the same dimension as the hidden latent image space. Then the concatenation of the projected text query and the hidden latent image $h_{z,l-1}$ outputted by the last block will pass K transformer blocks. In each transformer block, we propose the modality-specific feedforward neural network (FFN) for text and image. Besides, the text query will skip the cross-attention layer. Following the transformer blocks, we separate the concatenation between the text query and the image hidden, and project the text query back to the text embedding space using a fully-connected layer. Furthermore, we build a cross-block skip connection from the noised text query e_t with time condition to the hidden text query $h_{e,l-1}$ outputted by the last block. The hidden text query $h_{e,l-1}$ is initialized to zero.

3.4. Diffusion-base Unified Training

In this section, we introduce the training paradigm to unify diffusion-based image generation training and cross-modal alignment learning into a single framework. Algorithm 1 illustrates the whole training algorithm.

To alternatively feed the required inputs and the conditions when performing diffusion-based generation and alignment tasks, we introduce the masking mechanism to erase the unnecessary input. Note that the whole inputs’ space includes the latent image z outputted by the image encoder, text condition c and text query e outputted by the text encoder. For the *image generation task*, we mask the text query e , and feed the noisy latent image z_t with time step t and text condition c into the following UNet model. While for *image-text alignment learning*, to avoid leaking textual information, we mask the text condition c , then feed the noised text query e_t with timestep t and the latent image z as condition. Considering the image generation loss \mathcal{L}_{IG} from Eq. 4 and the diffusion-based image-text alignment loss \mathcal{L}_{ITA} from Eq. 8, the total loss of DiffDis \mathcal{L}_{Total} can

Algorithm 2 Text-conditional Image Generation Sampling.

Input: Text y , Text Encoder \mathcal{T} , Image Decoder \mathcal{D} , Timestep T , Noise Schedule $\{\beta_t\}_{t=1}^T$, Classifier-free guidance scale w .

- 1: $z_T \sim \mathcal{N}(0, I)$
 - 2: $c = \mathcal{T}(y)$
 - 3: $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{k=1}^t \alpha_k$
 - 4: **for** $t = T, \dots, 1$ **do** \triangleright For simplify, t stands for t_z .
 - 5: $\hat{e}_z = (1 + w)\Phi_u(z_t, t, c) - w\Phi_u(z_t, t)$
 - 6: $z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{z_t - \sqrt{1 - \bar{\alpha}_t}\hat{e}_z}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1 - \bar{\alpha}_{t-1}}\hat{e}_z$
 - 7: **end for**
 - 8: **return** $\mathcal{D}(z_0)$
-

Algorithm 3 Image-text Alignment Inference.

Input: Image x , Text y' of Downstream Task, Image Encoder \mathcal{V} , Text Encoder \mathcal{T} , Timestep T , Noise Schedule $\{\beta_t\}_{t=1}^T$, Classifier-free guidance scale w .

- 1: $e_T \sim \mathcal{N}(0, I)$
 - 2: $z = \mathcal{V}(x)$
 - 3: $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{k=1}^t \alpha_k$
 - 4: **for** $t = T, \dots, 1$ **do** \triangleright For simplify, t stands for t_e .
 - 5: $\hat{e}_0 = (1 + w)\Phi_\theta(e_t, t, z) - w\Phi_\theta(e_t, t)$
 - 6: $\hat{e}_e = (e_t - \sqrt{\bar{\alpha}_t}\hat{e}_0) / \sqrt{1 - \bar{\alpha}_t}$
 - 7: $e_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{e_t - \sqrt{1 - \bar{\alpha}_t}\hat{e}_e}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1 - \bar{\alpha}_{t-1}}\hat{e}_e$
 - 8: **end for**
 - 9: $e_0 = e_0 / \|e_0\|$ \triangleright L2 Normalize
 - 10: $e = \mathcal{T}(y')$ \triangleright Extract Text Embedding
 - 11: $e = e / \|e\|$ \triangleright L2 Normalize
 - 12: Perform the similarity $e_0^\top e$ on downstream tasks.
-

be calculated as:

$$\mathcal{L}_{Total} = \mathcal{L}_{IG} + \lambda \mathcal{L}_{ITA} \quad (9)$$

where λ denotes the weight factor.

Algorithm 2 and Algorithm 3 show the inference algorithm of sampling processes of text-conditional image generation and image-text alignment, respectively. During inference, different sampling acceleration techniques, such as classifier-free guidance sampling [20], can also be seamlessly integrated into our framework for both tasks.

4. Experiments

In this section, we first describe the detailed experiment settings (Sec. 4.1). Then we show the results on zero-shot image classification, image-text retrieval and text-to-image generation (Sec. 4.2). Finally, we conduct ablation studies on our DiffDis to validate the effectiveness of implementation designs (Sec. 4.3).

4.1. Experiment Setting

Model Architecture. To obtain better image synthesis performance, we initialize the autoencoder and UNet from Stable Diffusion-v1-1¹. The transformer in model Φ_θ is trained from scratch and contains 6 transformer blocks with 768 model width and 64-dim attention heads. The text encoder is initialized from CLIP-ViT-L/14 [42].

Experiment Details. We pre-train models on Conceptual Caption Dataset (CC3M) [47] to evaluate our DiffDis’s effectiveness. The resolutions of the original image and the latent image are set as $256 \times 256 \times 3$ and $32 \times 32 \times 4$ respectively. Following Stable Diffusion, the pixel values of the image are normalized to $[-1, 1]$ and the autoencoder and text encoder are frozen. We pre-train our DiffDis model for 6 epochs using the AdamW [33] optimizer with weight decay of $1e-4$. The batch size is set as 256 and the learning rate is set as $1e-5$ with 1000 steps linear warmup and kept unchanged until the training is finished. The new parameters introduced for the discriminative tasks use the learning rate of $1e-4$. We simply set λ to 1 and randomly drop 10% text condition for image generation and 10% image condition by zeroing the image for diffusion-based image-text alignment to enable the classifier-free guidance [20]. Exponential moving average (EMA) is applied every iteration and the decay coefficient is set as 0.9999. The EMA model is utilized to evaluate the performance of downstream tasks.

Evaluations. For *zero-shot image classification*, we evaluate our proposed DiffDis model on 12 classification datasets, i.e., CIFAR10 [25], CIFAR100 [25], Caltech101 [26], StanfordCars [24], Flowers102 [38], Food101 [3], SUN39 [2], Describable Textures Dataset (DTD) [8], Aircrafts [35], OxfordPets [40], EuroSAT [16], and ImageNet [10]. We adopt DDIM sampler [49] as described in Algorithm 3 to predict the text embedding and calculate the similarity between predicted text embedding and text embedding. During inference, we use 8 sampling steps and classifier-free guidance scale [20] of 3. Following CLIP [42], we ensemble the prompt templates to improve the zero-shot classification performance by averaging the text embeddings across different prompt templates.

For *zero-shot image-to-text retrieval* and *text-to-image retrieval tasks*, we conduct experiments on karpathy split test set [23] of MSCOCO [30] and Flickr30K [41] which are widely used benchmark datasets. The inference processing of retrieval is similar to image classification.

For *zero-shot text-to-image generation*, we evaluate the performance on 30,000 text prompts from MSCOCO dataset [30] under the evaluation of FID, KID by applying torch-fidelity [39]. We also calculate CLIP-Score [17] under pre-trained CLIP-RN50. We compare our proposed

¹Stable-Diffusion-v1-1 checkpoint: <https://huggingface.co/CompVis/stable-diffusion-v1-1-original>

	CIFAR10	CIFAR100	Caltech101	StanfordCars	Flowers102	Food101	SUN397	DTD	Aircrafts	OxfordPets	EuroSAT	ImageNet	Average
CLIP-ViT-B/32	63.0	28.5	58.2	1.1	12.4	12.8	24.3	7.6	1.5	11.5	11.8	16.7	20.8
CLIP-ViT-B/16	57.2	27.0	54.8	1.0	12.9	13.6	28.8	10.1	1.1	10.3	10.8	19.7	20.6
CLIP-ViT-L/14	57.5	28.1	57.2	1.8	10.8	14.3	31.2	12.1	1.7	11.7	24.9	21.1	22.7
DiffDis	57.1	32.3	68.6	3.1	16.9	16.2	35.1	26.4	2.0	28.5	17.5	25.9	27.4

Table 1. Top-1 accuracy(%) of zero-shot image classification on 12 datasets. Note that CLIP models are pre-trained on CC3M by using 4x larger batch size and 3.3x longer training epochs.

Model	Flickr30K						MSCOCO						Mean R@1
	image-to-text			text-to-image			image-to-text			text-to-image			
	R@1	R@5	R@10										
CLIP-ViT-B/32	18.8	42.3	53.9	12.5	30.5	39.9	10.1	25.3	35.3	7.0	18.9	27.1	12.1
CLIP-ViT-B/16	30.1	56.4	68.9	19.4	42.1	52.7	14.4	34.6	46.1	10.3	26.5	36.6	18.6
CLIP-ViT-L/14	29.9	58.3	70.4	20.3	46.1	57.6	14.7	35.0	47.1	11.3	28.4	38.9	19.1
DiffDis	49.8	77.5	85.6	38.8	67.9	77.6	26.3	50.9	62.8	19.5	42.2	54.2	33.6

Table 2. Results of zero-shot image-text retrieval on Flickr30K and MSCOCO datasets. ‘R@K’ means top-K recall. ‘Mean R@1’ means the average R@1 of image-to-text retrieval and text-to-image retrieval on Flickr30K and MSCOCO. Note that CLIP models are pre-trained on CC3M by using 4x larger batch size and 3.3x longer training epochs.

DiffDis with Stable Diffusion fine-tuned on CC3M with the same training hyper-parameters. All models use PNDM sampler [31] with 50 sampling steps under the classifier-free guidance scale [20] of 3.

4.2. Main Results

Zero-shot Image Classification. Table 1 presents detailed results on 12 classification datasets, comparing CLIP-ViT-B/32, CLIP-ViT-B/16, and CLIP-ViT-L/14 pre-trained on the same dataset, i.e., CC3M. To ensure a fair comparison with DiffDis, all CLIP models’ text encoders are initialized from the pre-training [42]. Despite CLIP models being pre-trained with a 4x larger batch size and 3.3x longer training epochs, DiffDis achieves an average accuracy gain of 4.7% across 12 datasets, outperforming CLIP-ViT-L/14. Moreover, DiffDis demonstrates significant performance improvements on some domain-specific datasets such as OxfordPets, which we attribute to its strong generation ability to capture fine-grained information.

Zero-shot Image-Text Retrieval. Table 2 presents experimental results on image-to-text (I2T) retrieval and text-to-image (T2I) retrieval for Flickr30K and MSCOCO datasets. We can observe that DiffDis outperforms CLIP-ViT-L/14 in R@1 of I2T retrieval and T2I retrieval by 11.6% and 8.2%, respectively, on the MSCOCO dataset. Moreover, on the Flickr30K dataset, DiffDis achieves superior results with 19.9% and 18.5% improvements in R@1 of I2T retrieval and T2I retrieval, respectively.

Zero-shot Text-to-Image Generation. In addition to the Stable Diffusion model pre-trained on CC3M, we com-

Model	Text-to-Image Generation)		
	FID↓	KID↓	CLIP-Score↑
OFA (Finetuned) [53]	10.5	-	-
LAFITE [58]	26.9	-	-
DALLE [44]	17.8	-	-
GLIDE [36]	12.2	-	-
DALLE2 [43]	10.4	-	-
Stable Diffusion	10.8	2.9e-3	24.6
DiffDis	9.8	2.3e-3	24.4

Table 3. Quantitative evaluation of FID and CLIP-score on MSCOCO dataset for zero-shot text-guided 256×256 image synthesis. The Stable Diffusion and DiffDis are pre-trained on CC3M.

Task	FID↓	ZS-Acc↑	Mean R@1↑
Dis	-	11.31	15.82
Gen	43.47	-	-
Hybrid	41.05	12.96	19.72

Table 4. Results of DiffDis with train-from-scratch UNet.

pare our proposed DiffDis model with LAFITE [58], DALLE [44], GLIDE [36], DALLE2 [43], and OFA [53]. Table 3 demonstrates that our proposed DiffDis model makes a comparable performance to Stable Diffusion, achieving a 1.0 improvement in FID and similarly CLIP-Score, indicating that dual-task learning helps in image generation tasks. Additionally, DiffDis’s zero-shot image generation performance on the MSCOCO dataset surpasses OFA, which has been fine-tuned on the MSCOCO dataset. Note that OFA is also a hybrid model to combine discrim-

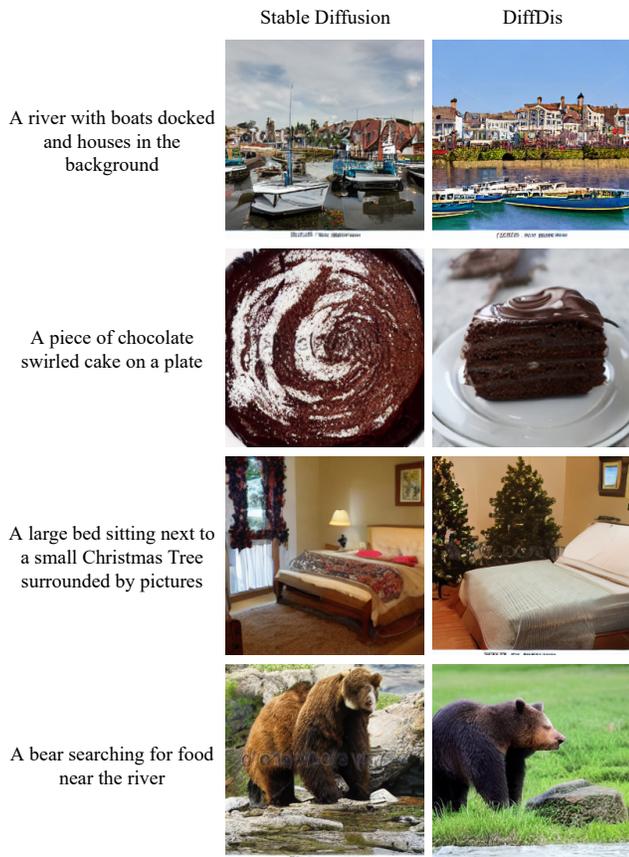


Figure 4. Qualitative comparisons of Stable diffusion and our DiffDis on MSCOCO zero-shot text-to-image generation.

inative tasks and generative tasks. Figure 4 illustrates the synthetic samples under the text condition generated by different models on MSCOCO.

4.3. Ablation Study

The Effect of Dual-task Learning. Table 4 presents a comparison between dual-task learning and single-task learning. In this experiment, UNet is trained from scratch, and the text encoder is frozen, which is the same as stable diffusion pre-training. Compared to single-task learning, dual-task learning shows a 1.65% improvement in zero-shot ImageNet classification, a 3.9% improvement on average R@1 of Flickr30k and MSCOCO, and a 2.42 improvement in FID of zero-shot MSCOCO text-to-image generation. These results demonstrate the effectiveness of unifying cross-modal generative and discriminative pre-training into a single framework under the diffusion process.

The Effect of Freezing Text Encoder. The generative task usually applies a frozen text encoder but the discriminative task often trains the text encoder. Table 5 demonstrates that freezing the text encoder obtains better performance on all

Freeze \mathcal{T}	Enlarge LR	FID↓	ZS-Acc↑	Mean R@1↑
✗	✓	11.56	25.72	29.37
✓	✗	10.98	24.98	31.08
✓	✓	9.80	25.92	33.60

Table 5. The effect of enlarging the learning rate for train-from-scratch parameters and freeze the text encoder \mathcal{T} .

γ	FID↓	ZS-Acc↑	Mean R@1↑
1	11.90	22.35	28.59
0.1	11.62	22.97	29.08
0.01	11.52	23.45	28.64

Table 6. Results of DiffDis with different scale factor γ for text embedding in diffusion-based image-text alignment. The model is trained without enlarging the learning rate, freeze \mathcal{T} , and deep fusion blocks.

Classifier-free Guidance Scale	None	2	3	4	5
ZS-Acc (ImageNet)↑	23.15	23.44	23.45	23.43	23.37
FID (MSCOCO)↓	30.13	12.59	11.52	13.25	15.18

Table 7. The ablation of classifier-free guidance level [20]. By reformulating in a diffusion framework, Classifier-free guidance can be applied to boost both generative and discriminative tasks.

tasks, especially in the image generation task. We freeze the text encoder in our main results.

Enlarging the Learning Rate of Train-From-Scratch Parameters. During pre-training, we utilize a learning rate for the train-from-scratch parameters, i.e., the additional parameters for discriminative tasks, that is 10 times larger than the base learning rate. Specifically, we set the learning rate of $1e-4$ for train-from-scratch parameters and the pre-trained parameters use a learning rate of $1e-5$. Table 5 demonstrates that enlarging the learning rate of train-from-scratch parameters can improve the performance on three downstream tasks.

The Effect of Text Embedding Scale Factor γ . The experimental results presented in Table 6 examine the impact of text embedding scale factor γ for discriminative learning on three downstream tasks. The results show that as γ decreases, the performance of text-guided image generation and zero-shot ImageNet classification improves.

The Effect of Classifier-free Guidance. Table 7 demonstrates that classifier-free guidance can enhance both image generation and discriminative ability. Specifically, the results indicate that the use of classifier-free guidance with a value of 3 leads to the best performance in terms of image synthesis and zero-shot ImageNet classification.

The Effect of Different Sampling Steps. In image generation tasks, larger steps for denoising typically result in better performance. This section aims to analyze the impact

Steps	1	4	8	10	20	50
ZS-Acc	14.85	23.14	23.15	23.12	23.13	23.13

Table 8. The performance on zero-shot ImageNet classification with different sampling steps. No classifier-free guidance.

Dataset	CIFAR10	CIFAR100	Food101	SUN397	ImageNet
Mean	57.15	32.31	16.25	35.15	25.92
Var	3.99e-05	1.66e-05	6.27e-06	8.88e-05	9.20e-05

Table 9. The stochasticity of DiffDis on zero-shot classification.

of various generation steps on the discriminative task. The experimental results are presented in Table 8. The results indicate that using 8 steps leads to the best performance on zero-shot ImageNet classification task. However, increasing the number of steps beyond 8 results in longer inference times without a significant improvement in performance.

Stochasticity of the Discriminative Evaluation. We run 10 times evaluations of zero-shot classification with different random seeds on five classification datasets to calculate the mean and variance of the performance. Table 9 shows that the discriminative performance is stabled.

5. Conclusion

This paper proposes a novel framework named DiffDis, which unifies the training of generative and discriminative tasks under the diffusion process. Initially, we formulate the image-text alignment into a diffusion process by adopting the text embeddings as diffusion objectives. We then propose a dual-stream network architecture that can simultaneously learn to reconstruct the image and text embeddings. As the first work to unify the training of generative and discriminative tasks under the diffusion process, we conducted a careful ablation study of the different key settings of the unified model. We observed that the unified model achieves improvement in both discriminative tasks and generative tasks compared to the single-task model. We hope that our work can serve as an exploration for future research on unifying discriminative tasks under the diffusion process, thereby providing more choices for future multi-task multi-modal joint-training model frameworks.

6. Acknowledgments

We gratefully acknowledge the support of MindSpore², CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

²<https://www.mindspore.cn/>

References

- [1] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khruikov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2021.
- [2] A. Barriuso and A. Torralba. Notes on image annotation. Preprint arXiv:1210.3448, 2012.
- [3] L. Bossard, M. Guillaumin, and L. Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- [4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, 2020.
- [5] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Computer Vision and Pattern Recognition*, 2021.
- [6] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- [8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Computer Vision and Pattern Recognition*, 2014.
- [9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.
- [10] J. Deng. A large-scale hierarchical image database. 2009.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.
- [12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [13] 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.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

- Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- [17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022.
- [20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.
- [22] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.
- [23] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [25] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [26] F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 2006.
- [27] L. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang. Visualbert: A simple and performant baseline for vision and language. Preprint arXiv:1908.03557, 2019.
- [28] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.
- [29] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. Preprint arXiv:2103.00823, 2021.
- [30] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.
- [31] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022.
- [32] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.
- [34] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Motlaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.
- [35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021.
- [38] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing*, 2008.
- [39] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738.
- [40] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *Computer Vision and Pattern Recognition*, 2012.
- [41] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision*, 2015.
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

- Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- [47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [50] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.
- [51] B. Thomee, D. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.
- [52] Changyao Tian, Wenhai Wang, Xizhou Zhu, Xiaogang Wang, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. *arXiv preprint arXiv:2111.13579*, 2021.
- [53] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.
- [54] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.
- [55] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.
- [56] Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *ArXiv*, abs/2208.07791, 2022.
- [57] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.
- [58] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.

A. Failure Case

The performance of image generation is relatively unsatisfactory (shown in Fig. 5) considering the following reasons. 1). Since we train the model on CC3M [47], which contains images of general scenes, the generation quality of some specific domains like humans, animals is low. Training data from these domains may further improve the generation quality (upper). 2). The generation results may contain watermarks since some images in CC3M are watermarked (bottom).



Figure 5. Failure cases of text-to-image generation.

B. More Implement Details

In this section, we introduce more implementation details for our DiffDis. 1) We use a cosine noise scheduler for the text query diffusion process and a linear noise scheduler for the image diffusion process. 2). We assign the timestep of 1000 to the image condition when performing discriminative tasks. Note that 1000 is not in the range of the timestep for image generation.

Here we give detailed experimental settings for the CLIP models we compared in the main paper. We set the batch size to 1024 and pre-training was conducted for 20 epochs by using AdamW optimizer. The learning rate is $1e-3$ and the weight decay is 0.1. During pre-training, the images are randomly cropped and we use the RandAugment [9] for image augmentation. We compare our implementation with open source clip pretraining codebase [21]. We keep the same batch size and the number of training epochs. The experimental results are shown in Table 10. Our implementation is better than open source codebase. We think that the improvement can be attributed to more extensive augmentation for images.

C. More Discussion

The Effect of Different Model Targets. Diffusion model’s output can be the original noise ϵ or the data x_0 that denote the noise prediction model and data prediction model, respectively. The comparison of two types of models on three downstream tasks is listed on Table 11.

Codebase	Model	ZS-Acc
OpenCLIP [21]	CLIP-ViT-B/32	14.7
OpenCLIP [21]	CLIP-ViT-L/14	19.1
Our	CLIP-ViT-B/32	16.7
Our	CLIP-ViT-L/14	21.1

Table 10. Comparison of our implementation and open source implementation [21].

Model Target	FID↓	ZS-Acc↑	Mean R@1↑
Noise	10.78	22.62	29.38
Data	11.52	23.44	28.64

Table 11. The performance of different model targets. Using feature scaling $\gamma = 1$.

Noise Scheduler	FID↓	ZS-Acc↑	Mean R@1↑
Cosine	11.90	22.35	28.59
Linear	11.52	22.70	31.07

Table 12. The performance of different noise schedulers. Using feature scaling $\gamma = 1$.

Enabled Fusion	FID↓	ZS-Acc↑	Mean R@1↑
✗	10.05	24.53	32.97
✓	9.80	25.92	33.60

Table 13. The performance on FID score on MSCOCO image generation, zero-shot ImageNet classification and average R@1 of MSCOCO and Flickr30k by enabling dual-stream deep fusion attention block.

The Effect of Different Noise Schedulers. We analyze the influence of different noise schedulers on the text diffusion process. The linear schedule starts from 0.00085 to 0.0120. Table 12 shows that the linear schedule is a better choice than the cosine schedule.

The Effect of Dual-Stream Deep Fusion Attention Block. To evaluate the effectiveness of the proposed dual-stream deep fusion attention block, we disable the fusion block by replacing it with the original attention blocks of Stable Diffusion. We directly concatenate the input text query with the image hidden output from UNet’s middle block and feed the concatenation to the 6 blocks transformer. Table 13 shows the experimental results of this comparison. When disabling the deep fusion block, the performances of three downstream tasks are dropped. Besides, according to Table 14, using modality-specific FFN and sharing the attention module in dual-stream deep fusion attention block will improve the performance on generation tasks.

Time Comparison. We provide the training time, generative inference time on COCO and discriminative inference time on ImageNet in Table 15. After unifying the discriminative and generative tasks, DiffDis has a longer train-

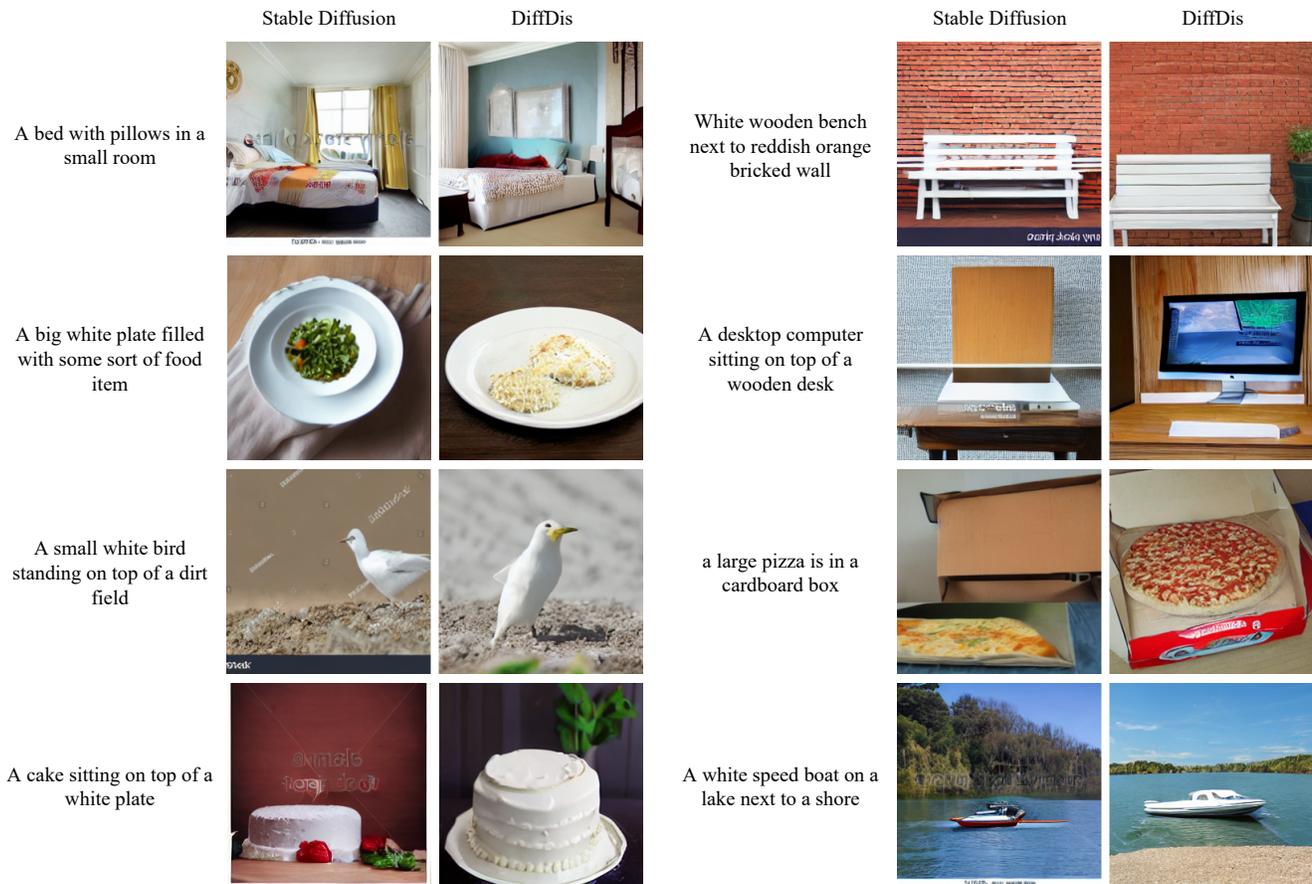


Figure 6. More illustrations of generated samples with proposed DiffDis on MSCOCO prompts.

Share Attn	MS-FFN	FID↓	ZS-Acc↑	Mean R@1↑
✓	✗	10.26	<u>25.92</u>	33.75
✗	✓	10.19	26.25	33.07
✓	✓	9.80	<u>25.92</u>	<u>33.60</u>

Table 14. The effect of the modality-specific FFN (MS-FFN) and sharing attention module in the dual-stream deep fusion attention block. We use the setting of the last row in our model.

ing time compared to single-task training but has a shorter training time than the sum training time of CLIP-ViT-L/14 and Stable Diffusion and make better or comparable performance. DiffDis has a similar generative inference time as Stable Diffusion and 1.7x discriminative inference time compared to CLIP.

The Mask Timestep of Image Condition for the Discriminative Tasks. The image condition for the discriminative tasks needs a timestep to input. We discuss the selection of the image condition on three downstream tasks on Table 16. The experimental results show that reusing the timestep

Time / Tasks	Training	Gen-Inference	Dis-Inference	ZS-Acc↑	FID↓
CLIP-ViT-L/14	1d 7h	–	148s	21.1	–
Stable Diffusion	1d 8h	3530s	–	–	10.8
DiffDis	2d 6h	3550s	252s	25.9	9.8

Table 15. The training time and inference time comparison.

within the range of image generation’s timestep leads to performance degradation on both image generation tasks and discriminative tasks. The use of the ‘First’ mask timestep ($t_z = 0$) will degrade the performance most. Assigning an additional timestep for the image condition for discriminative tasks achieves the best performance on all downstream tasks.

Discussion with HybViT We clarify that the DiffDis cannot directly compare with HybViT [56] since 1) HybViT focuses on class-condition image generation while our DiffDis targets text-condition image generation; 2) HybViT performs supervised classification tasks but can not perform zero-shot classification tasks or image-text retrieval tasks while DiffDis can.

Position	t_z	FID↓	ZS-Acc↑	Mean R@1↑
First	0	12.35	21.97	27.20
Last	999	12.02	21.73	27.56
Additional	1000	11.35	22.13	27.56

Table 16. Results of different mask timestep of image condition for discriminative learning. The range of the image generation diffusion steps is 0-999 . The additional timestep used for discriminative tasks is not shared with image generation.

Backbone	Pre-train Stage			Fine-tune Stage
	Image-Acc	Text-Acc	KNN-Acc	Acc
CLIP-ViT-L/14	31.4	38.1	35.5	40.5
DiffDis	37.0	52.5	40.5	44.4

Table 17. Results of long-tailed recognition on Places-LT dataset by using different backbone. We follow the official code of VL-LTR [52].

D. Application of DiffDis

We follow VL-LTR [52] to perform long-tailed visual recognition tasks and apply DiffDis or CLIP-ViT-L/14 (our implementation, pre-trained on CC3M), as the backbone. As shown in Table 17 We evaluate the performances of the pre-train stage and fine-tune stage on the Places-LT dataset [32].