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

Runhui Huang\textsuperscript{1} Jianhua Han\textsuperscript{2} Guansong Lu\textsuperscript{2} Xiaodan Liang\textsuperscript{1}\textsuperscript{*}
Yihan Zeng\textsuperscript{2} Wei Zhang\textsuperscript{2} Hang Xu\textsuperscript{2}
\textsuperscript{1}Shenzhen Campus of Sun Yat-sen University \textsuperscript{2}Huawei 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) \cite{17, 45, 51} such as Stable Diffusion \cite{42} and DallE2 \cite{40} have shown impressive results in image synthesis and re-define the capacity of state-of-the-art text-guided image synthesis. Typ-

\textsuperscript{*}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.](image)

\#15713
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) [14] 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 [52] tried to replace the U-Net structure in GLIDE [33] with a ViT [13] 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 [50], and Unified I/O [31] try to unify different vision-language tasks into an autoregressive sequence prediction framework. However, the autoregressive image generation solutions such as DallE [41] and OFA [50] show inferior performance compared to the DM-based models, such as DallE2 [40] and Stable Diffusion [42], 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 [10, 4, 12]. Recent works such as CLIP [39] and ALIGN [20] 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 [48], CC12M [5]). The VLP models can be categorized by their pre-training tasks: (a) Image-text contrastive learning task: CLIP [39], ALIGN [20], FILIP [53] and UNIMO [26] 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 \( e \) and text query \( q \) 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 \( e \) as the condition to predict the added noise \( \epsilon_e \). 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 \( \epsilon_e \).

### 3. Methodology

In this section, we first review some preliminaries about diffusion models (Sec. 3.1). Then we introduce our formulations of generative and discriminative 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 \( x_1, \ldots, x_T \) by progressively adding Gaussian noise to the sample according to some variance schedule given by \( \beta_t \) as follows:

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

and then learn a model to approximate the true posterior:

\[
p_\theta(x_{t-1} | x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),
\]

to perform the reverse denoising 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 \( \mathcal{L}_{V.L.B} \) on \( \log p_\theta(x_0) \) can be used to optimize \( \mu_\theta \) and \( \Sigma_\theta \), to achieve better results, Ho et al. [17] instead adopt a denoising 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:

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

where \( t \) uniformly sampled from \( \{1, \ldots, T\} \).

#### 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, \( \Phi_u \) and \( \Phi_t \) 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 [33, 18, 40, 43, 42], 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 \( \epsilon \) to gradually convert the original image \( x_0 \) to a random Gaussian distribution \( \mathcal{N}(z_T) \). Following the LDM [42], 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:

\[
\mathcal{L}_{IG} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), t \leq T} \left[ \| \epsilon - \Phi_u(z_t, t, z, c) \|^2 \right]
\]

(4)

where text condition \( c = T(y) \in \mathbb{R}^{L \times d_y} \) 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 \( \hat{x} = \mathcal{D}(z_0) \). DDIM sampler [46] and classifier-free guidance [19] 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 [53, 20]. The image-text contrastive (ITC) loss, like in CLIP [39], 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 \( s_{i,j}^x \) and \( s_{i,j}^y \) 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 [39].

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

\[
\begin{align*}
\hat{e}_t &= \gamma \sqrt{\alpha_t} e_0 + \sqrt{1-\alpha_t} \epsilon_t, \\
\alpha_t &= 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{j=0}^{t} \alpha_j,
\end{align*}
\]

(5)

(6)

where we omit the index \( i \), and \( \beta_t \) is used to control the strength of added noise for timestep \( t \). 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 \( \Phi_t \) is trained to estimate the original clean text query \( \hat{e}_0 = \Phi_t(\hat{e}_t, t, e) \). Note that \( \Phi_t \) can also be a noise prediction model to predict the noise \( \epsilon_e \).

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 \( e \):

\[
s = \hat{e}_0^\top e,
\]

(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:

\[
\begin{align*}
\mathcal{L}_{ITA} &= -\frac{1}{2B} \sum_{i=1}^{B} \left[ \log \sum_j \exp(s_{i,j}^x) + \log \sum_j \exp(s_{i,j}^y) \right],
\end{align*}
\]

(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 [42] 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,
we adopt the noisy latent image $z_t$ and text condition $c$ as input to predict the noise $e_z$ added on latent image $z$. For diffusion-based image-text alignment learning, we adopt the latent image $z$ and noisy text query $e_t$ as input to predict the original clean text query $e$. Note that it is viable to predict the noise $e_z$ added on text query $e$ for the alignment learning. More training details can refer to Sec. 3.4. Besides, the $\Phi(t)$ 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 [42] 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-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 $h_{e,l-1}$ 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 from 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 noisy text query $e_t$ with timestep $t$ and the latent image $z$ as condition. Considering the image generation loss $L_{IG}$ from Eq. 4 and the diffusion-based image-text alignment loss $L_{ITA}$ from Eq. 8, the total loss of DiffDis $L_{Total}$ can
be calculated as:

\[ L_{\text{Total}} = L_{IG} + \lambda L_{ITA} \]  

(9)

where \( \lambda \) 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 [19], can also be seamlessly integrated into our framework for both tasks.

### 4.1. Experiment Setting

**Model Architecture.** To obtain better image synthesis performance, we initialize the autoencoder and U-Net from Stable Diffusion-v1-1\(^1\). The transformer in model \( \Phi_T \) 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 [39].

**Experiment Details.** We pre-train models on Conceptual Caption Dataset (CC3M) [44] 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 [30] 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 \( \lambda \) 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 [19]. 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 [23], CIFAR100 [23], Caltech101 [24], StanfordCars [22], Flowers102 [35], Food101 [3], SUN39 [2], Describable Textures Dataset (DTD) [8], Aircrafts [32], OxfordPets [37], EuroSAT [15], and ImageNet [9]. We adopt DDIM sampler [46] 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 [19] of 3. Following CLIP [39], 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 [21] of MSCOCO [28] and Flickr30K [38] 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 [28] under the evaluation of FID, KID by applying torch-fidelity [36]. We also calculate CLIP-Score [16] under pre-trained CLIP-RN50. We compare our proposed

---

\(^1\)Stable-Diffusion-v1-1 checkpoint: https://huggingface.co/CompVis/stable-diffusion-v1-1-original
### 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 [39]. 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, 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 compare our proposed DiffDis model with LAFITE [54], DALLE [41], GLIDE [33], DALLE2 [40], and OFA [49]. 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-
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 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 $\gamma$. The experimental results presented in Table 6 examine the impact of text embedding scale factor $\gamma$ for discriminative learning on three downstream tasks. The results show that as $\gamma$ 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 generative 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.
We run Stochasticity of the Discriminative Evaluation. times without a significant improvement in performance. The number of steps beyond 8 results in longer inference times without a significant improvement in performance. However, increasing the number of steps beyond 8 results in longer inference times without a significant improvement in performance. Table 8. The performance on zero-shot ImageNet classification with different sampling steps. No classifier-free guidance.

<table>
<thead>
<tr>
<th>Steps</th>
<th>1</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>Food101</th>
<th>SUN397</th>
<th>Imagenet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>57.15</td>
<td>32.31</td>
<td>16.25</td>
<td>35.15</td>
<td>25.92</td>
</tr>
<tr>
<td>Var</td>
<td>3.99e-05</td>
<td>1.66e-05</td>
<td>6.27e-06</td>
<td>8.88e-05</td>
<td>9.20e-05</td>
</tr>
</tbody>
</table>

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 MindSpore®️️, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References


[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypymy whole, im-


