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

1. Failure Case

The performance of image generation is relatively unsatisfactory (shown in Fig. 1) considering the following reasons. 1). Since we train the model on CC3M [4], 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 1. Failure cases of text-to-image generation.

2. 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 [1] for image augmentation. We compare our implementation with open source clip pretraining codebase [2]. We keep the same batch size and the number of training epochs. The experimental results are shown in Table 1. Our implementation is better than open source codebase. We think that the improvement can be attributed to more extensive augmentation for images.

Codebase	Model	ZS-Acc
OpenCLIP [2]	CLIP-ViT-B/32	14.7
OpenCLIP [2]	CLIP-ViT-LB/14	19.1
Our	CLIP-ViT-B/32	16.7
Our	CLIP-ViT-L/14	21.1

Table 1. Comparison of our implementation and open source implementation [2].

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

Table 2. The performance of different model targets. Using feature scaling $\gamma = 1$.

3. 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 2.

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 3 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 4 shows the experimental results of this comparison. When disabling the deep fusion block, the performances of three

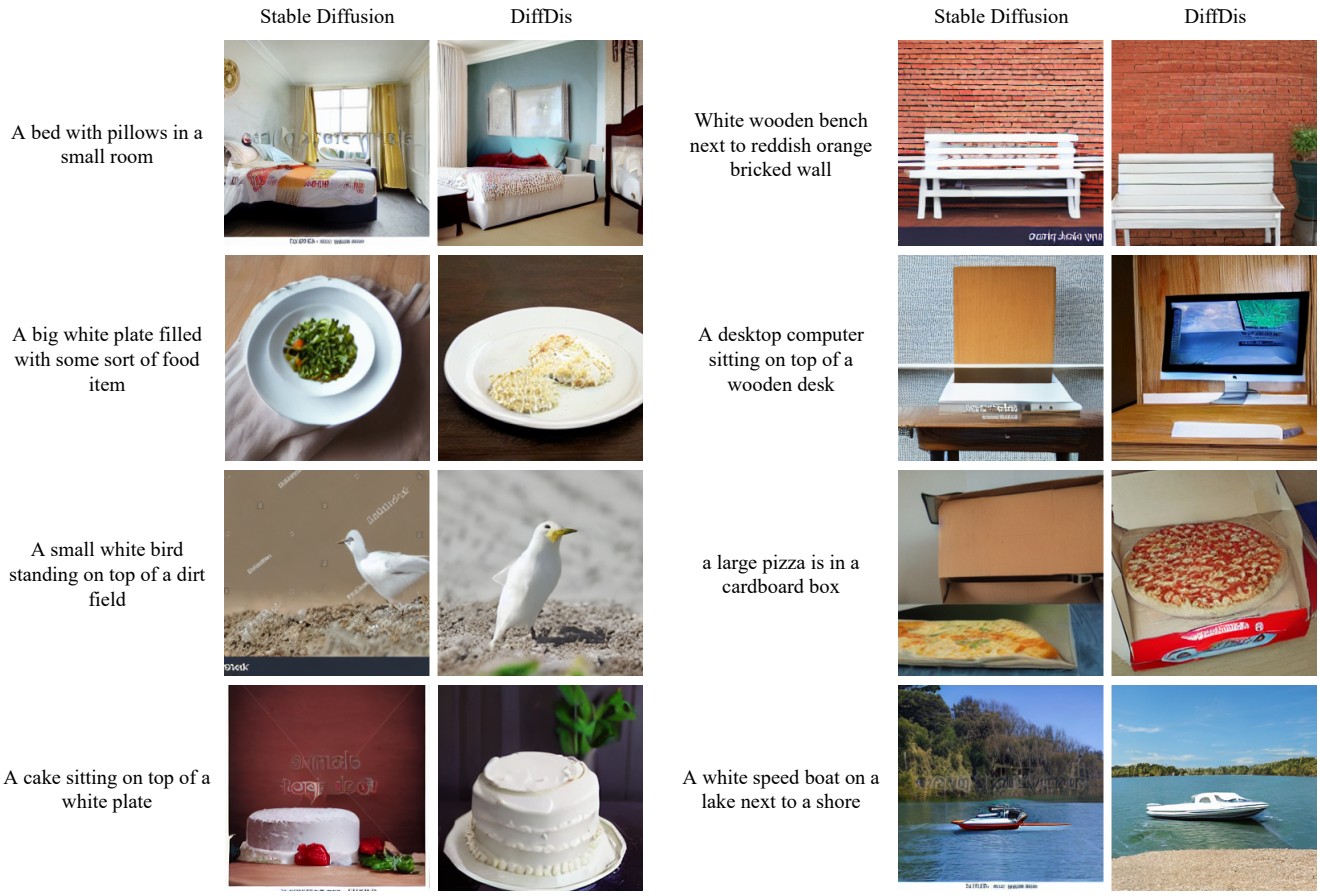

Figure 2. More illustrations of generated samples with proposed DiffDis on MSCOCO prompts.

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

Table 3. 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 4. 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.

downstream tasks are dropped. Besides, according to Table 5, 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 6. After unifying the discriminative and generative tasks, DiffDis has a longer training time compared to single-task training but has a shorter training time than the sum training time of CLIP-ViT-L/14 and Stable Diffu- sion 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.

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 5. 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.

The Mask Timestep of Image Condition for the Discriminative Tasks. The image condition for the discriminative

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 6. The training time and inference time comparison.

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 7. 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 8. Results of long-tailed recognition on Places-LT dataset by using different backbone. We follow the official code of VL-LTR [5].

tasks needs a timestep to input. We discuss the selection of the image condition on three downstream tasks on Table 7. The experimental results show that reusing the timestep 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 [6] 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.

4. Application of DiffDis

We follow VL-LTR [5] 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 8 We evaluate the performances of the pre-train stage and fine-tune stage on the Places-LT dataset [3].

References

[1] 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.

[2] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[3] 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.

[4] 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.

[5] 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.

[6] 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.