Supplementary Material for 'Free-ATM: Harnessing Free Attention Masks for Representation Learning on Diffusion-Generated Images'

David Junhao Zhang[1*], Mutian Xu[2,4†], Jay Zhangjie Wu[1], Chuhui Xue[3], Wenqing Zhang[3], Xiaoguang Han[2,4], Song Bai[3], and Mike Zheng Shou[1‡]

[1] Show Lab, National Univeristy of Singapore
[2] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[3] ByteDance
[4] The Future Network of Intelligence Institute, CUHK-Shenzhen

Section A More Related Work and Discussion

Representation Learning on Synthetic Data. Prominent studies such as BigdatasetGAN [10] and StableRep [18] have delved into the domain of representation learning on synthetic images. These endeavors, along with our own, can be delineated into two stages as shown in Fig. A. **1)**The first stage pertains to the creation of synthetic images. During this stage, BigdatasetGAN employs BigGAN for the production of synthetic images and further enhances this process by using human-annotated semantic masks to train a segmentation branch. This branch is subsequently utilized to derive the semantic masks for the synthetic images. StableRep adopts a diffusion model and employs captions for a single class to generate multiple samples within a positive pair. Differently, our approach utilizes diffusion models to generate both the images and the attention masks. This is advantageous as diffusion models naturally yield these attention masks during the image generation process, thereby obviating the need for additional computational efforts and the necessity for additional human annotation of masks.

2) As shown in Fig. A, the second stage is representation learning on synthetic images. BigdatasetGAN introduces an additional segmentation head within the contrastive learning framework to facilitate predictions and compute the loss between these predictions and the semantic masks generated in the first stage. StableRep employs multiple samples from one positive pair to refine contrastive learning. Our method, however, leverages the free attention masks and simple yet effective adaptation to augment various representation learning frameworks including vision-language pretraining, masked image modeling, and contrastive learning. After these stages, the pretrained backbones are applied to different downstream tasks.

* Work is partially done during an internship at ByteDance.
† Project lead.
‡ Corresponding author.

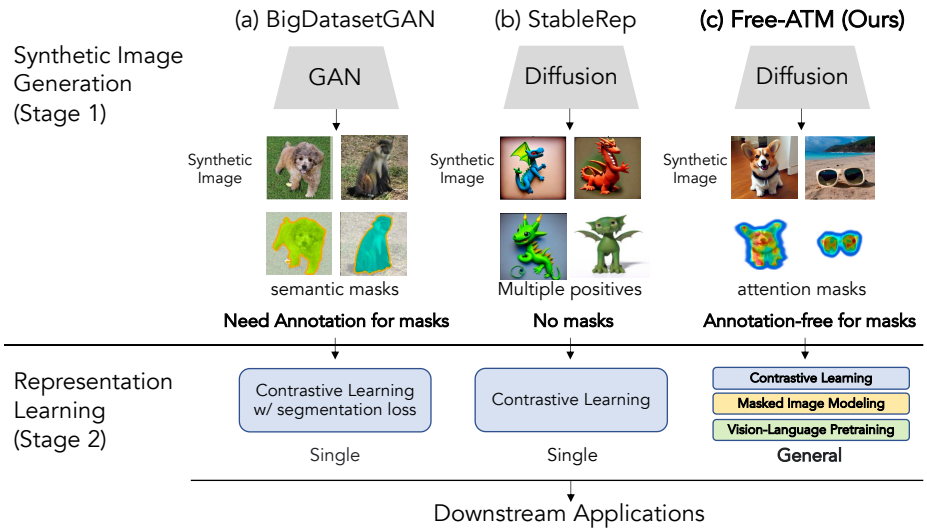

Fig. A: Overall Processes of representation learning on synthetic images and differences among BigdatasetGAN [10], StableRep [18] and our Free-ATM.

Table A: Comparisons with StableRep [18].

	PASVOC AP^b_{50}	COCO AP^b	Cityscapes $mIoU$
stable diffusion [17] synthetic images			
StableRep	79.3	36.5	73.9
Ours	**80.0**	**37.8**	**74.7**

Since StableRep also focuses on diffusion-generated data and contrastive learning, we draw comparisons with Free-ATM, as detailed in Table A. Both StableRep and our Free-ATM utilize SimCLR as the base contrastive learning framework for fair comparisons. The synthetic dataset was derived from Stable Diffusion 1.5, with the total number of images matching that of ImageNet-1K. As illustrated in Table A, our approach outperforms StableRep, demonstrating that our Free-ATM enhances contrastive learning on synthetic images through instance-level feature mining.

Attention Masks in Representation Learning. The works in [8, 11, 13, 14] implement strategies involving attention masks for representation learning. Specifically, [8, 13] employ an additional teacher model to generate attention masks in the pretraining stage. In contrast, [11, 14] develop intricate modules aimed at identifying semantic attention masks during pretraining stage. Our approach, however, diverges by generating attention masks at the synthetic image creation stage, allowing for their direct application in the pretraining stage without the need for elaborate mechanisms or extra teacher models.

Table B: Synthetic images from one-hot class conditioned diffusion model.

	PASVOC AP^b_{50}	COCO AP^b	Cityscapes $mIoU$
MoCo-v2	79.9	36.2	71.1
w/ ours	81.0 (↑1.1)	37.5 (↑1.3)	72.4(↑1.3)

Table C: Robustness to different text to image models.

	PASVOC AP^b_{50}	COCO AP^b	Cityscapes $mIoU$
glide [15] synthetic images			
MoCo-v2	81.0	37.1	72.9
w/ ours	82.0 (↑1.0)	38.9 (↑1.8)	74.1 (↑1.2)
stable diffsuion [17] synthetic images			
MoCo-v2	81.6	37.9	74.1
w/ ours	82.2 (↑0.6)	39.6 (↑1.7)	74.9 (↑0.8)

Table D: Robustness to different vision backbones.

	PASVOC AP^b_{50}	COCO AP^b	Cityscapes $mIoU$
RegNetX [15] synthetic images			
MoCo-v2	81.3	38.1	74.0
w/ ours	82.3 (↑1.0)	39.5 (↑1.4)	74.7 (↑0.7)
ResNet50 [5] synthetic images			
MoCo-v2	81.6	37.9	74.1
w/ ours	82.2 (↑0.6)	39.6 (↑1.7)	74.9 (↑0.8)

Section B More Ablation Studies

Different Condition for Diffusion Model. 1) In this paper, we aim to investigate the benefits of synthetic images for representation learning. Currently, text-based diffusion models, such as DALL-E 3, Deep-IF and Stable Diffusion, are the leading and most popular methods in image generation. Therefore, we use these text-based diffusion models, aligning with recent work [18]. **2)** Meanwhile, our method also works well on other conditions, e.g., the one-hot class label shown in Tab. B.

Robustness to Different Text-to-Image Models. In Table C, we assess the effectiveness of Free-ATM across various text-to-image diffusion models, including stable diffusion [17] and glide [15]. The findings indicate that Free-ATM enhances performance in both models, thereby evidencing the adaptability and strength of our approach with different text-to-image models.

Robustness to Different Vision Backbones. In our study on contrastive learning, as indicated in Table D, we investigate the robustness of Free-ATM across various vision backbones, specifically ResNet50 [5] and RegNetX [16]. In the context of vision-language pretraining, as presented in Table E, we examine the robustness of Free-ATM with different vision-language backbones, including Blip [12] and ViLT [9]. Our results demonstrate that Free-ATM improves perfor-

Table E: Robustness to different vision-language backbones.

	MS-COCO			
	finetune		zero-shot	
	tr@1	ir@1	tr@1	ir@1
ViLT [9]	41.1	38.6	20.3	17.5
w/ ours	44.8 (↑3.7)	41.8 (↑3.2)	28.8(↑8.5)	23.8(↑6.3)
Blip [12]	52.3	40.9	23.2	20.9
w/ ours	54.9 (↑2.6)	43.8 (↑2.9)	31.8(↑8.7)	28.2(↑7.3)

Table F: Robustness to the random seeds.

seeds	0-300	300-600	600-900
AP^b_{50}	82.33	82.24	82.21

Table G: Free-ATM on synthetic images containing single object.

	PASVOC AP^b_{50}	COCO AP^b	Cityscapes $mIoU$
MoCo-v2	81.1	37.3	73.4
w/ ours	81.8 (↑0.8)	38.9 (↑1.6)	74.2(↑0.8)

Table H: Comparisons with DetCon [6] and SEER [3] on synthetic images.

	PASVOC AP^b_{50}	COCO AP^b	Citysapces $mIoU$	training hours h
SEER	81.7	38.5	74.3	60
DetCon	81.9	38.8	74.3	81
Ours	**82.2**	**39.6**	**74.9**	**60**

mance across these backbones, highlighting its robustness in different vision and vision-language backbone scenarios.

Robustness to Different Seeds. We assess the stability of our approach to varying random seeds in the text-to-image model by using MOCO-V2 on the PASCAL VOC detection datasets. Table F illustrates that our method maintains robustness against the randomness introduced by different seeds in the text-to-image model.

Free-ATM on Synthetic Images Containing Single Object. As demonstrated in Table G, Free-ATM also enhances contrastive learning on synthetic images containing a single object. This effectiveness is attributed to how Free-ATM handles images with a single object. Without Free-ATM, formulating a positive pair is from two random crops of an image, which may represent the background and foreground areas respectively. During training, these areas, despite their differences, are brought closer together, leading ambiguity for learning discriminative features. Free-ATM can ensures that the features of both crops in a positive pair originate from the same instance, mitigating the aforementioned issue and improving pretraining outcomes

Comparisons with Object Prior Methods [6, 7]. We compare our method with other object prior methods on synthetic images. Object Prior Methods need

Table I: Comparisons with HPM [19] on synthetic images.

	ImageNet *acc*	ADE20K *miou*	Training memory G	Training hour hour
HPM	82.9	47.9	30.8G	298
Ours	**83.4**	**48.4**	**25.2G**	**264**

Table J: Using different mask strategies of MAE [4].

random	all high score masks	balance (ours)
82.7	82.7	83.4

to use non-trivial process to gradually get the object priors during the pre-training stage. For example, [6] uses a K-Means to determine the K instances masks in each training step, which needs multiple iterations and is quite time-consuming. In contrast, our approach leverages the Free-ATM of the diffusion mode, offering a direct and effective object localization without added steps in pretraining stage. In Table H, it is demonstrated that our approach yields superior outcomes and a faster pretraining process on synthetic data compared to DetCon [6], completing 200 training epochs on 16 V100 GPUs. These findings highlight the advantages of our methods when applied to synthetic images.

Comparisons with Multi-Crops Methods. We compare our method with the multi-crops method used in SEER [3], which employs 6 crops from each image following the SwAV [1]. Table H shows that MoCo-V2, when equipped with our Free-ATM, achieves better results than SEER in the context of pretraining on synthetic images.

Comparisons with HPM [19]. HPM introduces an additional teacher-student network to predict foreground patches. Meanwhile, the teacher network requires an extra learnable decoder with 8 transformer blocks for mask prediction during pre-training, increasing GPU memory and training duration. Instead, we do not need an additional teacher network and a heavy decoder. Our attention masks are directly from the diffusion model and can naturally fit in MAE [4] framework during pre-training w/o added training cost. As demonstrated in Table I of 1600 epochs training on 16 V100GPUs, our method outperforms HPM while incurring lower training costs.

Mask Strategies of MAE. As indicated in Section 3.3.2), we incrementally increase the proportion of masked patches chosen based on the highest importance scores as the number of training epochs increases. Concurrently, we decrease the proportion of masked patches selected at random. We test the effectiveness of this method by comparing it to an approach where all masked patches are determined solely by the highest importance scores. The downstream results on ImageNet classification [2] are presented in Table J, which suggest that balance masking proves to be more beneficial, yielding better performance.

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020) 5
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 5
3. Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al.: Self-supervised pretraining of visual features in the wild. arXiv preprint arXiv:2103.01988 (2021) 4, 5
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022) 5
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 3
6. Hénaff, O.J., Koppula, S., Alayrac, J.B., Van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: ICCV (2021) 4, 5
7. Hénaff, O.J., Koppula, S., Shelhamer, E., Zoran, D., Jaegle, A., Zisserman, A., Carreira, J., Arandjelović, R.: Object discovery and representation networks. In: ECCV (2022) 4
8. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. In: ECCV (2022) 2
9. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML (2021) 3, 4
10. Li, D., Ling, H., Kim, S.W., Kreis, K., Barriuso, A., Fidler, S., Torralba, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In: CVPR (2022) 1, 2
11. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. NeurIPS (2022) 2
12. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022) 3, 4
13. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., et al.: Mst: Masked self-supervised transformer for visual representation. NeurIPS (2021) 2
14. Liu, Z., Gui, J., Luo, H.: Good helper is around you: Attention-driven masked image modeling. In: AAAI (2023) 2
15. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022) 3
16. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollar, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 3
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 3
18. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. NeurIPS (2023) 1, 2, 3
19. Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., Zhang, Z.: Hard patches mining for masked image modeling. In: CVPR (2023) 5