Supplementary for DIFFGUARD: Semantic Mismatch-Guided Out-of-Distribution Detection using Pre-trained Diffusion Models

A. Implementation Details of DIFFGUARD

Pre-trained Weights of Diffusion Models. On CIFAR-10, we use the same pre-trained model as DiffNB[1], which is a conditional DDPM [2] with classifier-free guidance. On IMAGENET, we use the unconditional Guided Diffusion Model (GDM) [1] and apply classifier guidance[2]. While for Latent Diffusion Model (LDM) [5], we use the model pre-trained on IMAGENET[3] with classifier-free guidance.

Similarity Metric. To measure the similarity of image synthesis and the input, we adopt several similarity metrics [3], together with the logits from the classifier-under-protection as a commonly considered measure for out-of-distribution (OOD) detection [4]. Table A shows the metrics we consider for different benchmarks.

In general, we find that DISTS performs well on IMAGENET. Compared to other low-level metrics (e.g., ℓ_2), DISTS provides more robust image-space comparisons. For instance, if the generated image displays different brightness levels from the input, DISTS can offer a more consistent comparison than ℓ_2. This is also evidenced by LDM, where DISTS consistently outperforms ℓ_2. By contrast, since many similarity metrics (e.g., DISTS, LPIPS) apply pre-trained weights on IMAGENET as the feature extractor, they may not be suitable for CIFAR-10 directly. Thus, the logits distance works best on CIFAR-10. This result also enables a direct comparison between our DIFFGUARD and DiffNB [4] (in Table 1), where logits are also utilized as the distance metric.

It is important to note that in the main paper, we report the result only with one generic metric on different benchmarks, without combining different similarity metrics. In practice, it is feasible to combine multiple metrics for judgment. Such a combination can be either the one employed in Sec. 4.2 and Sec. 4.3, where distinct metrics are treated as additional baselines; or the one presented in [7], where various metrics are taken into account, and the rejection of OOD is based on any of them (*i.e.*, work in a tandem manner for OOD rejection).

[1]https://github.com/luping-liu/DiffOOD
[2]https://github.com/openai/guided-diffusion
[3]https://github.com/CompVis/latent-diffusion

benchmark	model	metrics	DDIM steps
CIFAR-10	DDIM	logits	50
IMAGENET	GDM	DISTS	100
IMAGENET	LDM	DISTS	25

Table A. Detailed settings of DIFFGUARD in the main paper for the different benchmarks, including similarity metrics and DDIM timesteps.

DDIM timesteps. In Table A, we present the DDIM timesteps utilized in Table 1 and Table 2 of the main paper. Specifically, for CIFAR-10, we opt for the same settings as DiffNB [4], using DDIM-50. According to Sec. 4.4, LDM is preferable for fewer DDIM timesteps, resulting in faster inference. In comparison, GDM typically performs better with more DDIM timesteps. To balance the speed and OOD detection performance, we adopt DDIM-100 in the main paper.

method	OOD dataset	AUROC ↑	FPR@95 ↓
GDM(oracle)	Species	87.35	54.97
	iNaturalist	94.15	31.60
	OpenImage-O	90.97	45.94
	ImageNet-O	86.22	62.20
	average	89.67	48.67
LDM(oracle)	Species	97.38	14.41
	iNaturalist	97.76	12.71
	OpenImage-O	95.12	25.12
	ImageNet-O	95.97	22.60
	average	96.56	18.71

Table B. Results for applying the oracle classifier with DIFFGUARD on the IMAGENET benchmark.

B. Use of the Oracle Classifier on IMAGENET

In Sec. 4.2, we presented the performance of DIFFGUARD on the CIFAR-10 benchmark using an oracle classifier. In this section, we demonstrate how DIFFGUARD performs on the IMAGENET benchmark with the help of an oracle classifier, as shown in Table B. We utilized the same settings as in Table 2. Our results indicate that the perfor-

mance of LDM and GDM can be significantly improved with an oracle classifier. Since the oracle classifier only provides the predicted label, while GDM relies on the gradient from the classifier, we resort to classifier-under-protection for gradient (*i.e.*, ResNet50). Therefore, its performance may be limited by the incorrect gradient estimation from the classifier. On the other hand, LDM employs classifier-free guidance, and therefore, both AUROC and FPR@95 demonstrate a significant improvement.

C. More Qualitative Results

Fig. B and Fig. C display the image syntheses by DIFFGUARD with LDM and GDM, respectively. The visualization reveals that DIFFGUARD can effectively produce analogous images in InD scenarios, while emphasizing the semantic mismatch in OOD scenarios. Regarding the comparison between GDM and LDM, we notice that GDM occasionally incorporates unrealistic features from the classifier-under-protection in the synthesized images, while LDM consistently generates photo-realistic syntheses, even in OOD cases. Such a phenomenon on GDM motivates us to employ adaptive early-stop (AES) in Sec. 3.2.1, Tech #2. Despite that LDM sometimes alters InD samples, GDM does not. This further justifies DIFFGUARD to extract and use the information from the classifier-under-protection for LDM, as stated in Sec. 3.2.2, Tech #3. As shown in Fig. B, only certain details are modified after applying DIFFGUARD with Tech #3, while the main structure and content are preserved.

D. Failure Case Analysis

We present some failure cases of DIFFGUARD in Fig. A. For InDs, these failures are mainly due to image synthesis problems. For example, we observe some test cases exhibit different fields of view from common cases, which makes DIFFGUARD difficult to maintain their original content. Additionally, certain classes (e.g. jellyfish and front curtain) tend to be monochrome or dark, which could cause generative models to fail in synthesizing such images. To address these issues, better generative models may help.

Regarding OODs, the major problem is that some cases appear visually similar to InD classes, or the area of semantic mismatch is limited. It is worth noting that DIFFGUARD can successfully depict the target semantics in these cases, but the image synthesis still looks similar to the input, resulting in difficulty to detect them by similarity measurements. To solve such cases, one possible solution is to utilize better similarity metrics for detailed comparisons (e.g. feature distance from a model trained with contrastive learning [6]).

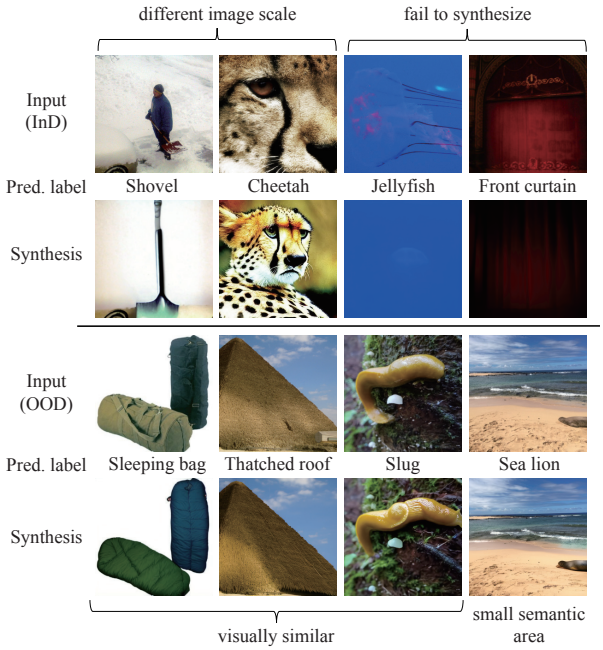

Figure A. Some failure cases for DIFFGUARD. DIFFGUARD may fail when image synthesis fails for InDs or when the content of OODs is indeed visually similar to InDs' semantics.

References

[1] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8780–8794, 2021. 1

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020. 1

[3] Sergey Kastryulin, Jamil Zakirov, Denis Prokopenko, and Dmitry V. Dylov. Pytorch image quality: Metrics for image quality assessment, 2022. 1

[4] Luping Liu, Yi Ren, Xize Cheng, and Zhou Zhao. Out-of-distribution detection with diffusion-based neighborhood, 2023. 1

[5] 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 (CVPR)*, pages 10684–10695, 2022. 1

[6] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 2

[7] Yijun Yang, Ruiyuan Gao, and Qiang Xu. Out-of-distribution detection with semantic mismatch under masking. In *European Conference on Computer Vision (ECCV)*, pages 373–390. Springer, 2022. 1

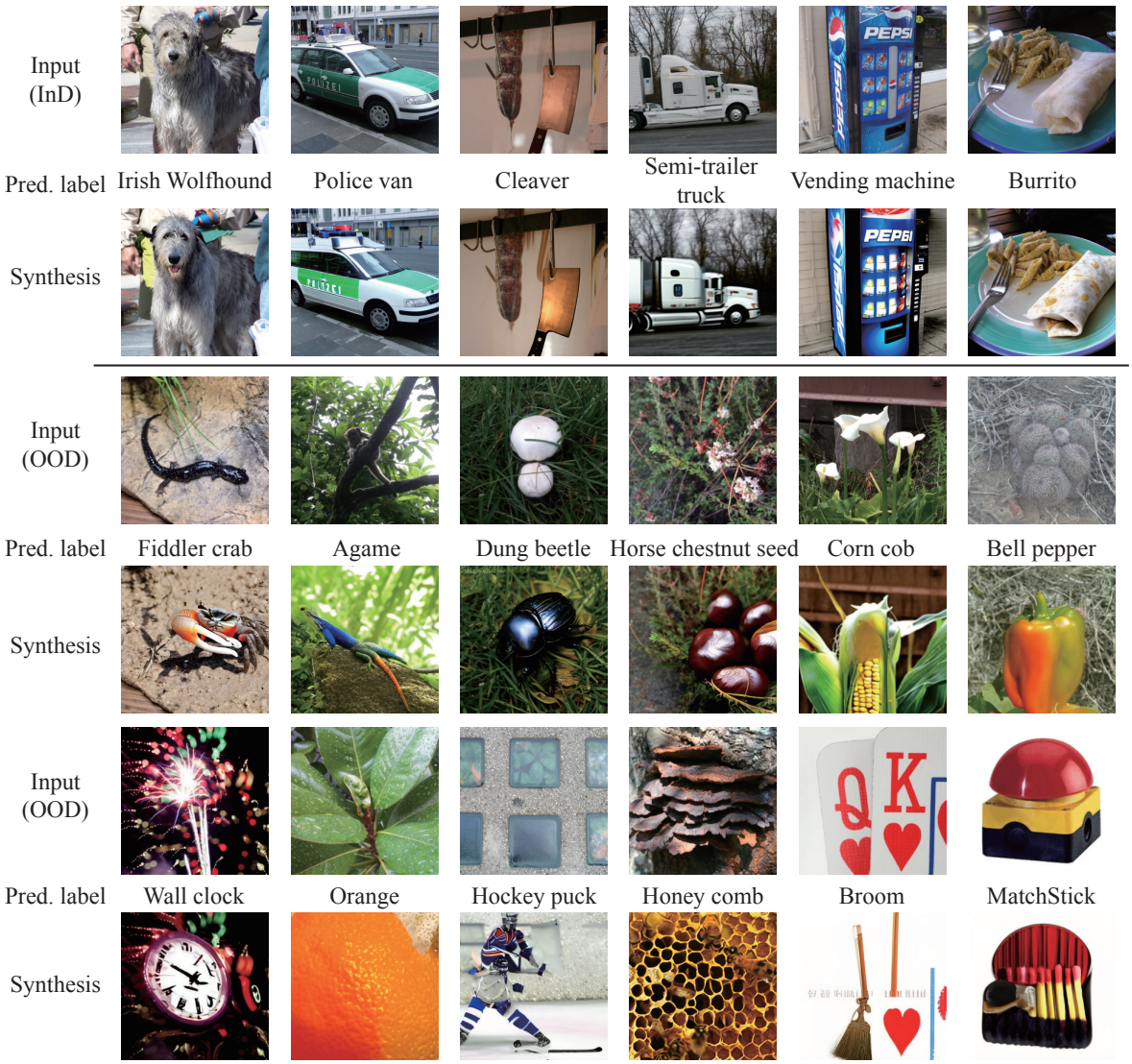

Figure B. Visualization for InD and OOD cases with their syntheses according to the predicted labels. Images are from the IMAGENET benchmark. We use LDM in this figure, *i.e.* classifier-free guided diffusion. We can identify a clear similarity difference between InDs and OODs by comparing the inputs with their syntheses.

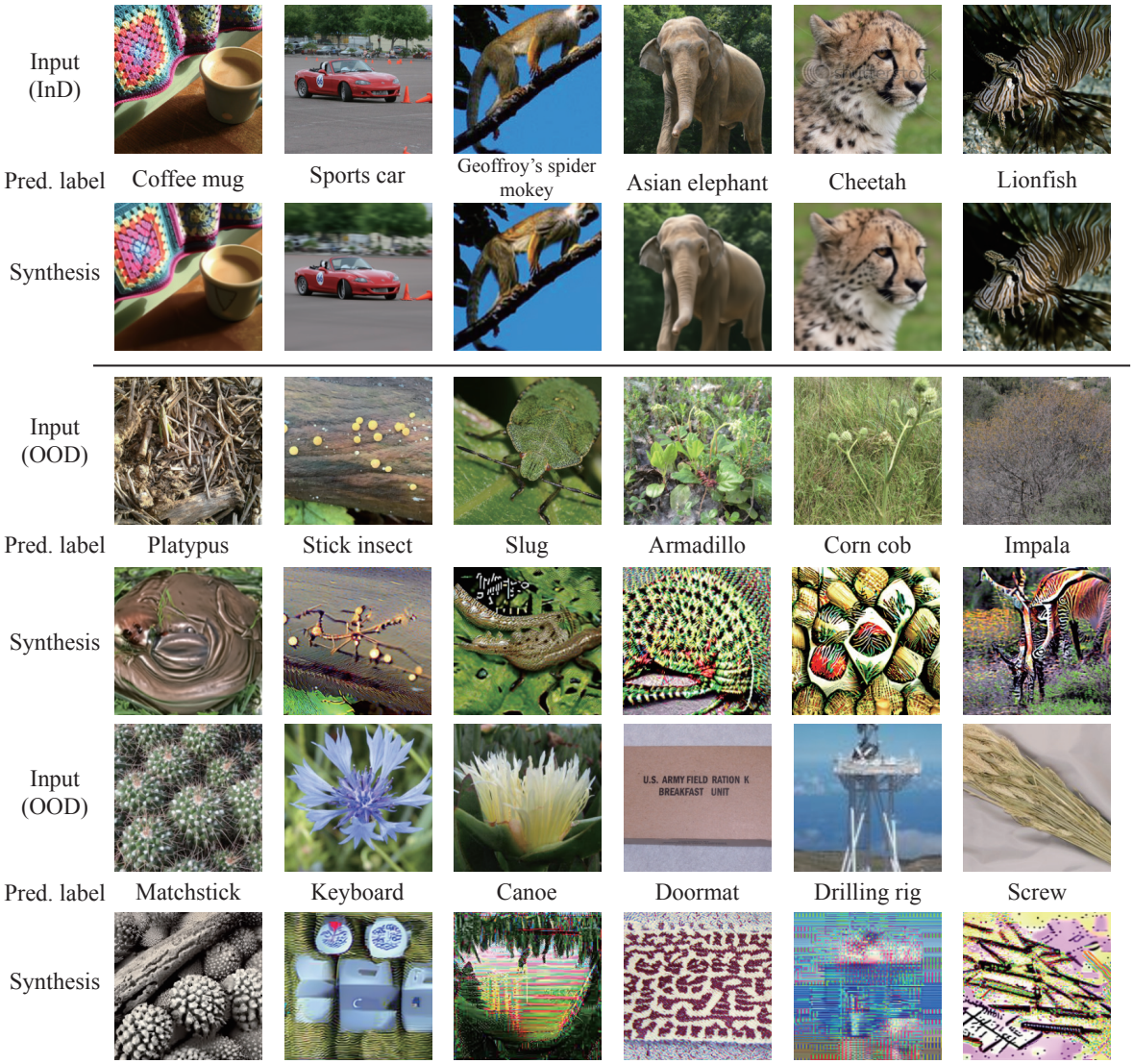

Figure C. Visualization for InD and OOD cases with their syntheses according to the predicted labels. Images are from the IMAGENET benchmark. We use GDM in this figure, *i.e.* classifier-guided diffusion. We can identify a clear similarity difference between InDs and OODs by comparing the inputs with their syntheses.