# Supplementary Material: Intriguing Properties of Diffusion Models: An Empirical Study of the Natural Attack Capability in Text-to-Image Generative Models

Takami Sato†[1], Justin Yue†[1], Nanze Chen†[2], Ningfei Wang[1], Qi Alfred Chen[1]
[1]University of California, Irvine
[2]University of Cambridge
{takamis, jpyue, ningfei.wang, alfchen}@uci.edu, nc630@cam.ac.uk

## 1. Detailed Overview of NDDA Dataset

The latest NDDA dataset consists of the following 15 classes: stop sign, car, dog, hot dog, traffic light, zebra, fire hydrant, frog, horse, bird, boat, air plane, bicycle, cat, and carrot with 6 diffusion models (Dall-E 2 [5], Dall-E 3 [6], Stable Diffusion 2 [15], Deepfloyd IF [16], Stable Diffusion 1.5 [15], MidJourney [2], and Google Duet [7]). The examples of the images generated by each diffusion model are shown in Fig. 7. For future versions of the dataset, we plan to generate additional variations of the prompts for each subject to capture a greater variety of NDD attacks. At the time of writing, the following 5 classes have a second variation: stop signs, horses, fire hydrants, cars, and cats. As shown in Fig. 1, the dataset is organized into 6 "diffusion parent folders" that separate each diffusion model's set of images, which in turn contains multiple folders for each of the 15 object classes from COCO. Each object class folder then contains multiple subfolders that hold the NDD attack images, and these subfolders' names are the text prompts used to generate the set of NDD attack images. For example, if images were generated using Stable Diffusion 2 with the text prompt, "blue dog", the path to this prompt subfolder would be "ndda_dataset/stable_diffusion_2/blue_dog/". For the purposes of submission, we downsample the images to a quarter of their original dimensions and only include 1 image per prompt subfolder.

```
NDDA_dataset
├── Dall-E_2
│   └── stop_sign
│       ├── stop_sign
│       │   ├── img1.jpg
│       │   ├── img2.jpg
│       │   └── ...
│       └── blue_stop_sign
│           └── ..
├── Midjourney
└── ..
```

Figure 1. Overview of the NDDA dataset directory structure, using the Dall-E folder as the example diffusion folder. For each object class folder (stop sign in this case), there are multiple prompt folders that contain ≥ 50 images for models with API access or ≥ 20 images for models w/o API access.

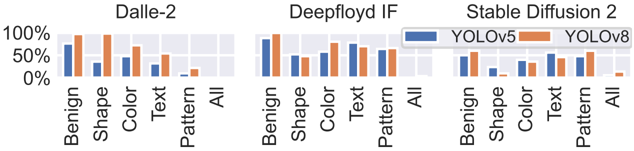


Figure 2. Detection rates of YOLOv5 and YOLOv8 on the stop sign images generated by the 3 diffusion models.

## 2. Additional Results of Natural Attack Capability on Object Detectors

Table 1 and 2 show the detection results for the fire hydrant and horse classes in the NDDA dataset generated by 3 diffusion models. The results of the stop sign are shown in the main paper. As shown, the majority of the images are still detected as keeping the targeted objects. Even if all robust features are removed, the 3 diffusion models are always able to generate effective attacks for 3 models other than YOLOv3 and YOLOv5.

### 2.1. Additional Evaluation with YOLOv8

We evaluate the natural attack capability against YOLOv8 [11], which is one of the current state-of-the-art object detectors. Fig. 2 illustrates the comparison with YOLOv5 in the stop sign case. As shown, YOLOv8 is generally more vulnerable to the NDD attack than YOLOv5, particularly for the images of DALL-E2. It thus indicates that the current state-of-the-art model still has vulnerability against the NDD attack.

†denotes co-first authors

Table 1. Detection rates of 5 object detectors on the **fire hydrant** images in the NDDA dataset generated by the 3 diffusion models. **Bold** and underline denote highest and lowest scores in each row.

| | Removed Robust Features | | | Object Detectors | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Shape | Color | Pattern | YOLOv3 | YOLOv5 | DETR | Faster | RTMDet | Avg. |
| DALL-E 2 | | | | 96% | 34% | **100%** | 98% | **100%** | 86% |
| | ✓ | | | 18% | 0% | 8% | 4% | **40%** | 14% |
| | | ✓ | | 98% | 38% | **100%** | 96% | **100%** | 86% |
| | | | ✓ | 58% | 4% | 88% | 84% | **92%** | 65% |
| | ✓ | ✓ | ✓ | 0% | 0% | 4% | 2% | **16%** | 4% |
| Stable Diffusion 2 | | | | 94% | 96% | **100%** | **100%** | 20% | 82% |
| | ✓ | | | 6% | 0% | 14% | **20%** | **20%** | 12% |
| | | ✓ | | 92% | 86% | 94% | **100%** | **100%** | 94% |
| | | | ✓ | 94% | 82% | **100%** | 98% | 98% | 94% |
| | ✓ | ✓ | ✓ | 0% | 0% | 4% | **6%** | 4% | 3% |
| Deepfloyd IF | | | | 98% | 84% | **100%** | **100%** | **100%** | 96% |
| | ✓ | | | 34% | 10% | 52% | 76% | **86%** | 52% |
| | | ✓ | | **98%** | 46% | **98%** | **98%** | **98%** | 88% |
| | | | ✓ | 96% | 52% | **100%** | 98% | 98% | 88% |
| | ✓ | ✓ | ✓ | **72%** | 8% | 7% | 66% | 84% | 47% |

Table 2. Detection rates of 5 object detectors on the **horse** images in the NDDA dataset generated by the 3 diffusion models. **Bold** and underline denote highest and lowest scores in each row.

| | Removed Robust Features | | | Object Detectors | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Shape | Color | Pattern | YOLOv3 | YOLOv5 | DETR | Faster | RTMDet | Avg. |
| DALL-E 2 | | | | 76% | 48% | 78% | 84% | **86%** | 74% |
| | ✓ | | | 90% | 60% | 92% | **96%** | 94% | 86% |
| | | ✓ | | **48%** | 4% | 40% | 32% | **48%** | 34% |
| | | | ✓ | 8% | 0% | 18% | 16% | **24%** | 13% |
| | ✓ | ✓ | ✓ | 10% | 0% | 14% | 2% | **18%** | 9% |
| Stable Diffusion 2 | | | | 86% | 60% | 90% | 88% | **94%** | 84% |
| | ✓ | | | **100%** | 92% | 98% | **100%** | **100%** | 98% |
| | | ✓ | | **82%** | 18% | 72% | 54% | 68% | 59% |
| | | | ✓ | 46% | 10% | 64% | 58% | **74%** | 50% |
| | ✓ | ✓ | ✓ | 68% | 12% | 50% | 60% | **74%** | 53% |
| Deepfloyd IF | | | | 96% | 54% | **98%** | 94% | 100% | 88% |
| | ✓ | | | 90% | 76% | 92% | **96%** | **96%** | 90% |
| | | ✓ | | 82% | 32% | 72% | 68% | **88%** | 68% |
| | | | ✓ | 60% | 2% | 64% | 62% | **70%** | 52% |
| | ✓ | ✓ | ✓ | **44%** | 2% | 28% | 14% | **44%** | 26% |

## 2.2. Additional Evaluation on More Class Categories

Fig. 3 shows box plots of the detection rates of the 15 object categories (stop sign, car, dog, hot dog, traffic light, zebra, fire hydrant, frog, horse, bird, boat, air plane, bicycle, cat, and carrot). We evaluate 6 object detectors: YOLOv3, YOLOv5, DETR, Faster R-CNN, RTMDet, and YOLOv8. Thus, there are 90 (15 × 6) data points for each plot. As shown, there are always some levels of vulnerability against the NDD attack because the median values are above zero. Our paper covered the general classes of the object: a stop sign for an artificial sign, a fire hydrant for an artificial object, and a horse for a natural object. We thus leave a deep analysis of other categories for the following works.

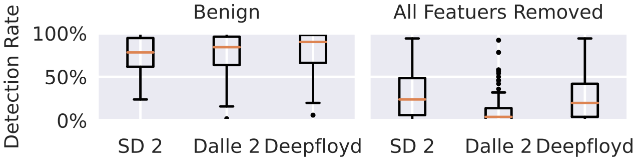


Figure 3. Box plots of detection rates for 15 object category images generated by the 3 diffusion models.

## 3. Detailed Results of Natural Attack Capability on Image Classification Models

*Experimental setup.* We train 4 state-of-the-art image classification models, ResNet50 [8], DenseNet121 [9], EfficientNet b0 [17], and ResNeXt [18], on the training datasets derived from the COCO dataset [13]. We convert the COCO dataset into a dataset for the multiclass classification task in the same way as we did in RQ4: We crop the images in the COCO dataset with their bounding box annotations and randomly select 500 images for each class.

*Results.* Table 3, 4, and 5 show the classification accuracy for the 3 classes: stop signs, fire hydrants, and horses. As shown, the large number of the generated images are still classified as the targeted object class even though we remove a robust feature. For stop sign, ≥47% of the generated images are still classified as stop signs even though all 4 robust features are removed. For fire hydrant and horse, ≥38% and ≥16% of the generated images are classified as the original class, respectively. In summary, the NDD attack shows quite high attack effectiveness against not only object detectors but also image classifiers.

## 4. Detailed Results of GAN-based Attacks (RQ1)

Fig. 4 shows the attacks generated by BigSleep that successfully trick object detectors with a confidence score ≥ 0.5. As shown, GAN-based NDD attacks can still generate several successful attacks which have high stealthiness as confirmed by the user study. While the diffusion models have much higher attack effectiveness as discussed in RQ1, this natural attack capability in the GAN model can be also a serious attack threat. Considering such a high attack capability has not been reported even though many GAN-based adversarial attacks [4] are proposed, we think this is mainly due to the high-quality text guide by OpenAI CLIP [14] in BigSleep [3] rather than due to a unique characteristic of

Table 3. Classification accuracy of 4 classifiers on the **stop sign** images in the NDDA dataset generated by the 3 diffusion models. **Bold** and underline denote highest and lowest scores in each row.

| | Removed Robust Features | | | | Image Classifiers | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Shape | Color | Text | Pattern | Resnet50 | Dense121 | Effic.Net | Resnext | Avg. |
| DALL-E 2 | | | | | **100%** | **100%** | **100%** | **100%** | 100% |
| | ✓ | | | | **98%** | **98%** | **98%** | **98%** | 98% |
| | | ✓ | | | **100%** | **100%** | **100%** | **100%** | 100% |
| | | | ✓ | | 94% | 94% | 94% | 96% | 95% |
| | | | | ✓ | 84% | 88% | 86% | **90%** | 87% |
| | ✓ | ✓ | ✓ | ✓ | 42% | 58% | 40% | **64%** | 51% |
| Stable Diffusion 2 | | | | | 72% | **76%** | 60% | 66% | 69% |
| | ✓ | | | | 50% | 56% | 40% | **80%** | 57% |
| | | ✓ | | | **68%** | 58% | 54% | **68%** | 62% |
| | | | ✓ | | 72% | 66% | 38% | **76%** | 63% |
| | | | | ✓ | **58%** | 56% | 44% | **58%** | 54% |
| | ✓ | ✓ | ✓ | ✓ | **56%** | 44% | 34% | 54% | 47% |
| Deepfloyd IF | | | | | **100%** | 96% | **100%** | **100%** | 99% |
| | ✓ | | | | **96%** | 70% | 84% | 92% | 86% |
| | | ✓ | | | **100%** | 86% | **100%** | **100%** | 97% |
| | | | ✓ | | 92% | 86% | 74% | **96%** | 87% |
| | | | | ✓ | 88% | 86% | 78% | **92%** | 86% |
| | ✓ | ✓ | ✓ | ✓ | **90%** | 86% | 52% | **90%** | 78% |

Table 4. Classification accuracy of 4 classifiers on the **fire hydrant** images in the NDDA dataset generated by the 3 diffusion models. **Bold** and underline denote highest and lowest scores in each row.

| | Removed Robust Features | | | Image Classifiers | | | | |
|---|---|---|---|---|---|---|---|---|
| | Shape | Color | Pattern | Resnet50 | Dense121 | Effic.Net | Resnext | Avg. |
| DALL-E 2 | | | | 96% | **98%** | 84% | 96% | 94% |
| | ✓ | | | **94%** | 86% | 78% | 76% | 84% |
| | | ✓ | | **94%** | 88% | 90% | 90% | 91% |
| | | | ✓ | **76%** | 62% | 36% | 52% | 57% |
| | ✓ | ✓ | ✓ | **66%** | 62% | 64% | 46% | 60% |
| Stable Diffusion 2 | | | | **90%** | **90%** | 78% | 84% | 86% |
| | ✓ | | | 50% | 46% | 40% | **52%** | 47% |
| | | ✓ | | **94%** | 86% | 88% | 86% | 89% |
| | | | ✓ | 68% | 64% | 62% | **72%** | 67% |
| | ✓ | ✓ | ✓ | **46%** | 26% | 34% | **46%** | 38% |
| Deepfloyd IF | | | | 98% | 98% | **100%** | **100%** | 99% |
| | ✓ | | | **94%** | 90% | **94%** | **94%** | 93% |
| | | ✓ | | 94% | 92% | 96% | **100%** | 96% |
| | | | ✓ | **96%** | 88% | 82% | 90% | 89% |
| | ✓ | ✓ | ✓ | 86% | 82% | 74% | **98%** | 85% |

Table 5. Classification accuracy of 4 classifiers on the **horse** images in the NDDA dataset generated by the 3 diffusion models. **Bold** and underline denote highest and lowest scores in each row.

| | Removed Robust Features | | | Image Classifiers | | | | |
|---|---|---|---|---|---|---|---|---|
| | Shape | Color | Pattern | Resnet50 | Dense121 | Effic.Net | Resnext | Avg. |
| DALL-E 2 | | | | 64% | 60% | 48% | **66%** | 60% |
| | ✓ | | | 80% | **84%** | 74% | 82% | 80% |
| | | ✓ | | 34% | 32% | 18% | **38%** | 31% |
| | | | ✓ | **18%** | 6% | 0% | 6% | 8% |
| | ✓ | ✓ | ✓ | **26%** | 14% | 10% | 12% | 16% |
| Stable Diffusion 2 | | | | **90%** | 78% | 74% | 82% | 81% |
| | ✓ | | | 92% | **98%** | 86% | 94% | 93% |
| | | ✓ | | **80%** | 54% | 40% | 54% | 57% |
| | | | ✓ | **32%** | 24% | 14% | 28% | 25% |
| | ✓ | ✓ | ✓ | 46% | **62%** | 48% | 54% | 53% |
| Deepfloyd IF | | | | **94%** | 92% | 82% | **94%** | 61% |
| | ✓ | | | 90% | **92%** | 86% | 82% | 88% |
| | | ✓ | | **94%** | 90% | 80% | 86% | 88% |
| | | | ✓ | 60% | **64%** | 44% | 50% | 55% |
| | ✓ | ✓ | ✓ | 52% | **62%** | 34% | 34% | 46% |

the GAN model. However, it is not trivial to investigate the relationship between the impact of non-robust features and the quality of the text guide. We hope that a future study can handle this. So far, the DDA should be more preferable to the attackers because BigSleep's image generation takes much longer (≥20 minutes) than the diffusion model (a few seconds), even though BigSleep needs to generate more images due to its low attack success rate.

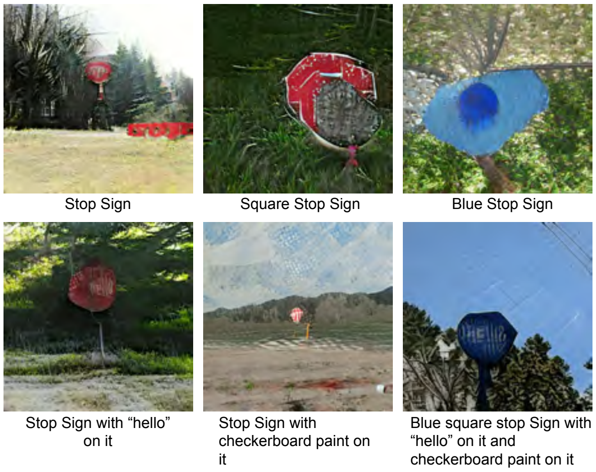

Stop Sign
Square Stop Sign
Blue Stop Sign
Stop Sign with "hello" on it
Stop Sign with checkerboard paint on it
Blue square stop Sign with "hello" on it and checkerboard paint on it

Figure 4. Examples of stop sign images generated by BigSleep that successfully trick at least one object detector, i.e., a stop sign is detected with a confidence score $\geq 0.5$. The user survey includes these images and more.

## 5. Detailed Setup of the User Study (RQ2)

To conduct an ethical user study, we have gone through the IRB process in our institution. The entire questionnaire form of this user study is attached as a part of the supplementary materials. We recruited 82 human subjects on Prolific [1], a crowdsourcing platform specialized for research purposes. All human subjects are English speakers in the United States to follow the IRB instruction. We did not collect any other demographic or privacy-sensitive information from the participants.

We provide 3 types of images: benign, diffusion-generated, and GAN-generated. 3 benign images are provided first in order to have a set of baseline results. Then, we show diffusion-generated images to the users; for every text prompt used, 3 images are generated. A text prompt may either produce a benign image, an image with 1 robust feature removed, or an image with all robust features removed. This step is also repeated using the BigSleep [3] model. Fig. 4 shows examples of stop sign images generated by BigSleep that successfully trick at least one object detector, i.e., a stop sign is detected with a confidence score ≥0.5. The user survey includes these images and more.

## 6. Additional Results of the Experiment on Non-Robust Features (RQ4)

Table 6 and 7 show the accuracy of the robust and normal classifiers on the fire hydrant and horse object classes, respectively. For the fire hydrant images, the robustified classifier drops the accuracy when the robust features are removed, as observed in RQ4. For the horse images, this analysis does not show meaningful results because the training of robustified classifier fails as the accuracy is <40% even for the benign images. We manually check the robustified horse images and find that the robustified images do not look like legitimate horses. Since this methodology [10] is originally evaluated on the CIFAR-10 [12], it may not be fully compatible with the dataset derived from the COCO dataset. Nevertheless, our finding is generally observed in other cases when the robustified classifiers can achieve similar accuracy to the normal classifier in the benign images.

Table 6. Accuracy of the robust and normal classifiers on the **fire hydrant** images in the NDDA dataset. The benign means the images generated with the benign prompts; The NDD means the NDD attack that removes all robust features.

| | Robustified classifier | | | Normal classifier | | |
|---|---|---|---|---|---|---|
| | Benign | NDD | Diff. | Benign | NDD | Diff. |
| DALL-E 2 | 0.66 | 0.1 | 0.56 | 0.94 | 0.36 | 0.58 |
| Stable Diffusion 2 | 0.96 | 0.42 | 0.54 | 0.96 | 0.38 | 0.58 |
| DeepFloyd IF | 0.82 | 0.2 | 0.62 | 1.00 | 0.9 | 0.1 |
| Avg. | 0.81 | 0.24 | **0.57** | 0.97 | 0.55 | **0.39** |

## 7. Additional Results of Tesla Experiments (RQ6)

Fig. 5 and 6 show the stop signs that successfully and unsuccessfully deceived Tesla's vision system respectively. We not only demonstrate that 73% of these generated images are successful but also show 3/4 of the diffusion models exhibit the natural attack capability in the real world. Out of the models that were successful, most of them required a carefully designed text prompt in order to achieve a successful attack, as illustrated by the last 5 images of Fig. 5 for Adobe Firefly and Stable Diffusion 2. Other diffusion models, such as Google Duet, require less effort; a simple text prompt, "stop sign", is enough to generate images that do not look like stop signs but have enough non-robust features to fool object detectors. We plan to inform this vulnerability to Tesla if this paper is accepted. We hope that our study can facilitate further research to train more robust DNN models.

Table 7. Accuracy of the robust and normal classifiers on the **horse** images in the NDDA dataset. The benign means the images generated with the benign prompts; The NDD means the attack that removes all robust features. Robustified classifier fails to train on the robustified data as its detection rate is ≤40% on the benign images

| | Robustified classifier | | | Normal classifier | | |
|---|---|---|---|---|---|---|
| | Benign | NDD | Diff. | Benign | NDD | Diff. |
| DALL-E 2 | 0.10 | 0.06 | 0.04 | 0.54 | 0.04 | 0.50 |
| Stable Diffusion 2 | 0.28 | 0.28 | 0.00 | 0.64 | 0.24 | 0.40 |
| DeepFloyd IF | 0.40 | 0.28 | 0.12 | 0.9 | 0.28 | 0.62 |
| Avg. | 0.26 | 0.21 | **0.05** | 0.69 | 0.19 | **0.51** |

## References

[1] Prolific. https://researcher-help.prolific.co/hc/en-gb, 2014. 4

[2] MidJourney. https://www.midjourney.com/, 2023. 1

[3] Adverb. BigSleep. https://github.com/lucidrains/big-sleep, 2021. 2, 4

[4] Tao Bai, Jun Zhao, Jinlin Zhu, Shoudong Han, Jiefeng Chen, Bo Li, and Alex Kot. AI-GAN: Attack-Inspired Generation of Adversarial Examples. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2543–2547. IEEE, 2021. 2

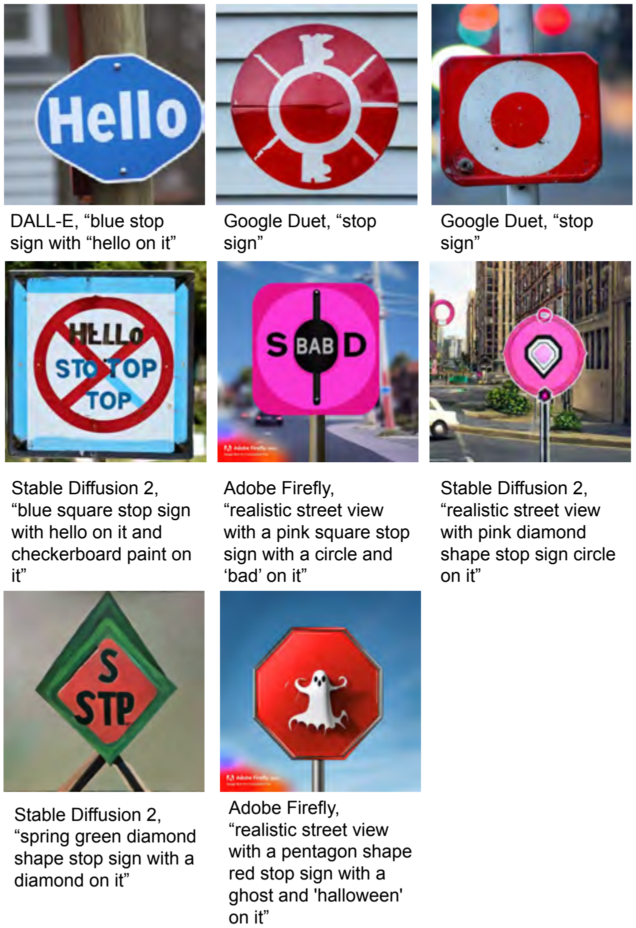
[5] DALL-E-2. DALL-E-2. https://openai.com/dall-e-2, 2022. 1

[6] DALL-E-3. DALL-E-3. https://openai.com/dall-e-3, 2023. 1

[7] Google. Introducing Duet AI for Google Workspace. https://workspace.google.com/blog/product-announcements/duet-ai, 2023. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 2



DALL-E, “blue stop sign with “hello on it”

Google Duet, “stop sign”

Google Duet, “stop sign”

Stable Diffusion 2, “blue square stop sign with hello on it and checkerboard paint on it”

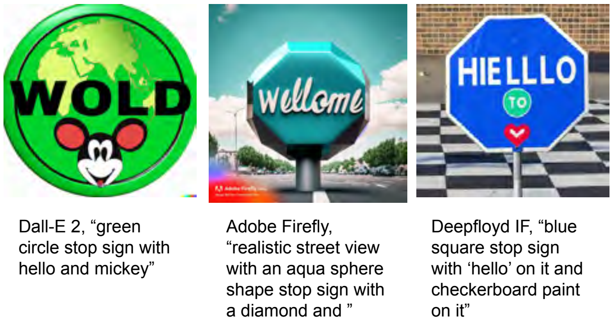
Adobe Firefly, “realistic street view with a pink square stop sign with a circle and ‘bad’ on it”

Stable Diffusion 2, “realistic street view with pink diamond shape stop sign circle on it”

Stable Diffusion 2, “spring green diamond shape stop sign with a diamond on it”

Adobe Firefly, “realistic street view with a pentagon shape red stop sign with a ghost and 'halloween' on it”

Figure 5. Successful NDD Attacks on Tesla Model 3. The caption of each image shows the used diffusion model and the text prompt.



Dall-E 2, “green circle stop sign with hello and mickey”

Adobe Firefly, “realistic street view with an aqua sphere shape stop sign with a diamond and ”

Deepfloyd IF, “blue square stop sign with ‘hello’ on it and checkerboard paint on it”

Figure 6. Unsuccessful NDD Attacks on Tesla Model 3. The caption of each image shows the used diffusion model and the text prompt.

[10] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 4

[11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO. `https://github.com/ultralytics/ultralytics`, 2023. 1

[12] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2009. 4

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 2, 6

[14] 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 (ICML)*, pages 8748–8763, 2021. 2

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[16] StabilityAI. DeepFloyd IF. `https://github.com/deep-floyd/IF`, 2023. 1

[17] Mingxing Tan and Quoc Le. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019. 2

[18] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017. 2

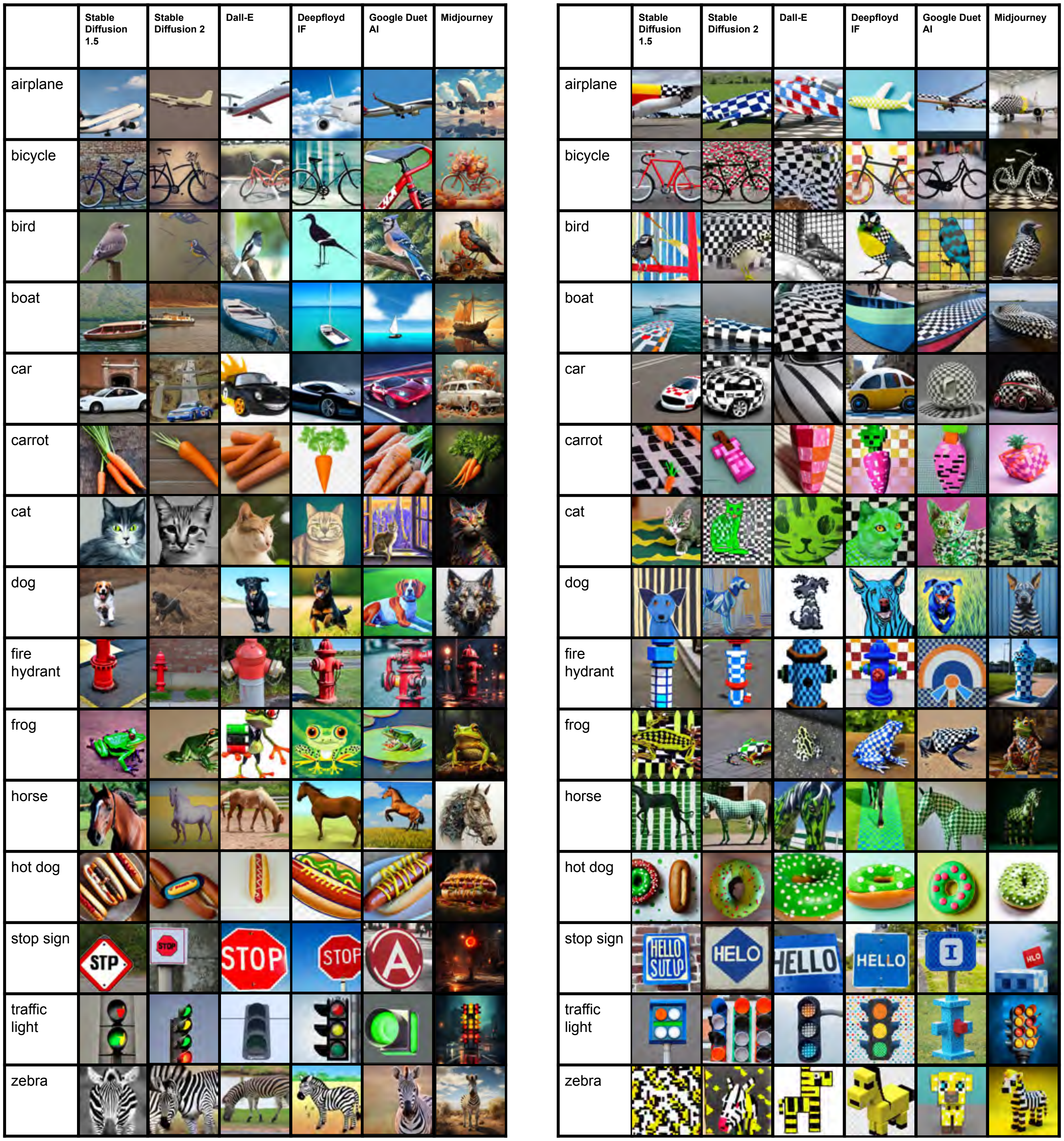


Figure 7. Overview of the NDDA dataset. 6 popular text-to-image diffusion models are used to generate 15 object classes from the COCO dataset [13]. The left grid consists of benign images while the right grid shows NDD attacked images where diffusion models are instructed to remove all robust features from the image.