

GENDATAAGENT: ON-THE-FLY DATASET AUGMENTATION WITH SYNTHETIC DATA

A MORE LLAMA-2 CAPTION PERTURBATION COMPARISON

The reason to combine "A photo of [Classname]" in the prompt is that captioning model BLIP-2 may fail to recognize the fine-grained categories and Figure 1 is one example. It can be observed that the identical caption with different seeds leads to similar generative images in guidance 7.5. Although guidance 2.0 introduces more diversity, the quality of generative images drops significantly. In comparison, using Llama-2 caption perturbation with guidance 7.5 can obtain high-quality yet diverse synthetic images.

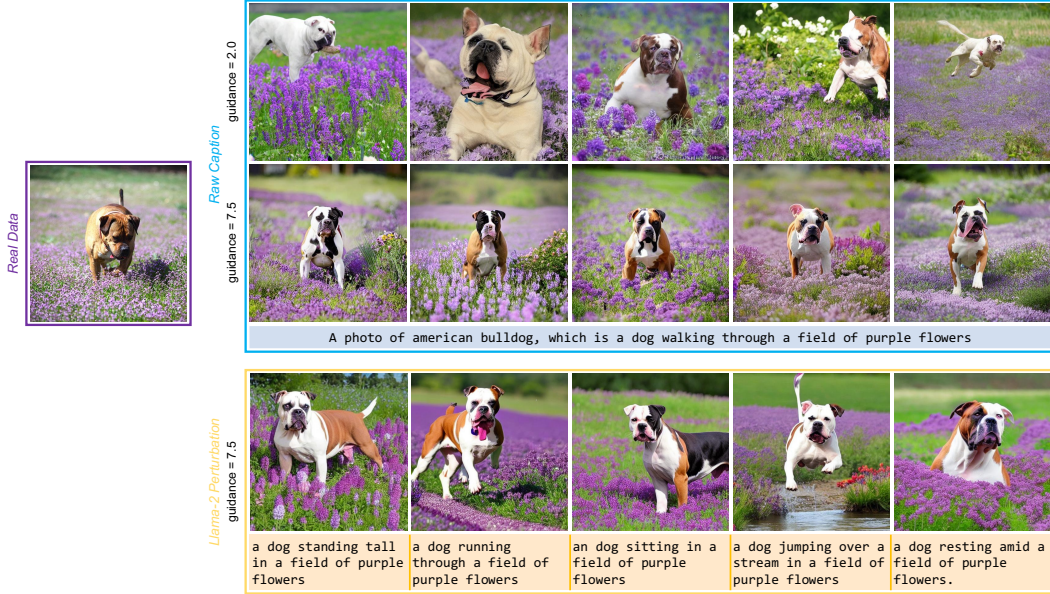


Figure 1: Comparison between raw caption with guidance 2.0 and guidance 7.5, as well as our Llama caption perturbation with guidance 7.5. For simplicity, the perturbed captions omit the prefix "A photo of [classname]".

B MORE VOG FILTERING EXAMPLES

We show more results of in-distribution synthetic data and outliers for all datasets in Figure 2.

C MULTIPLE RUNS

We conduct experiments with 3 different random seeds for synthetic data augmentation in Table 1.

D THRESHOLD ANALYSIS

Due to the page limit, we put some experiments and analyses in supplementary.

Image Strength. Following Real-Fake (Yuan et al., 2023), we combine the latent prior of real images to generate synthetic data, with image strength as the threshold for noise added to the reference

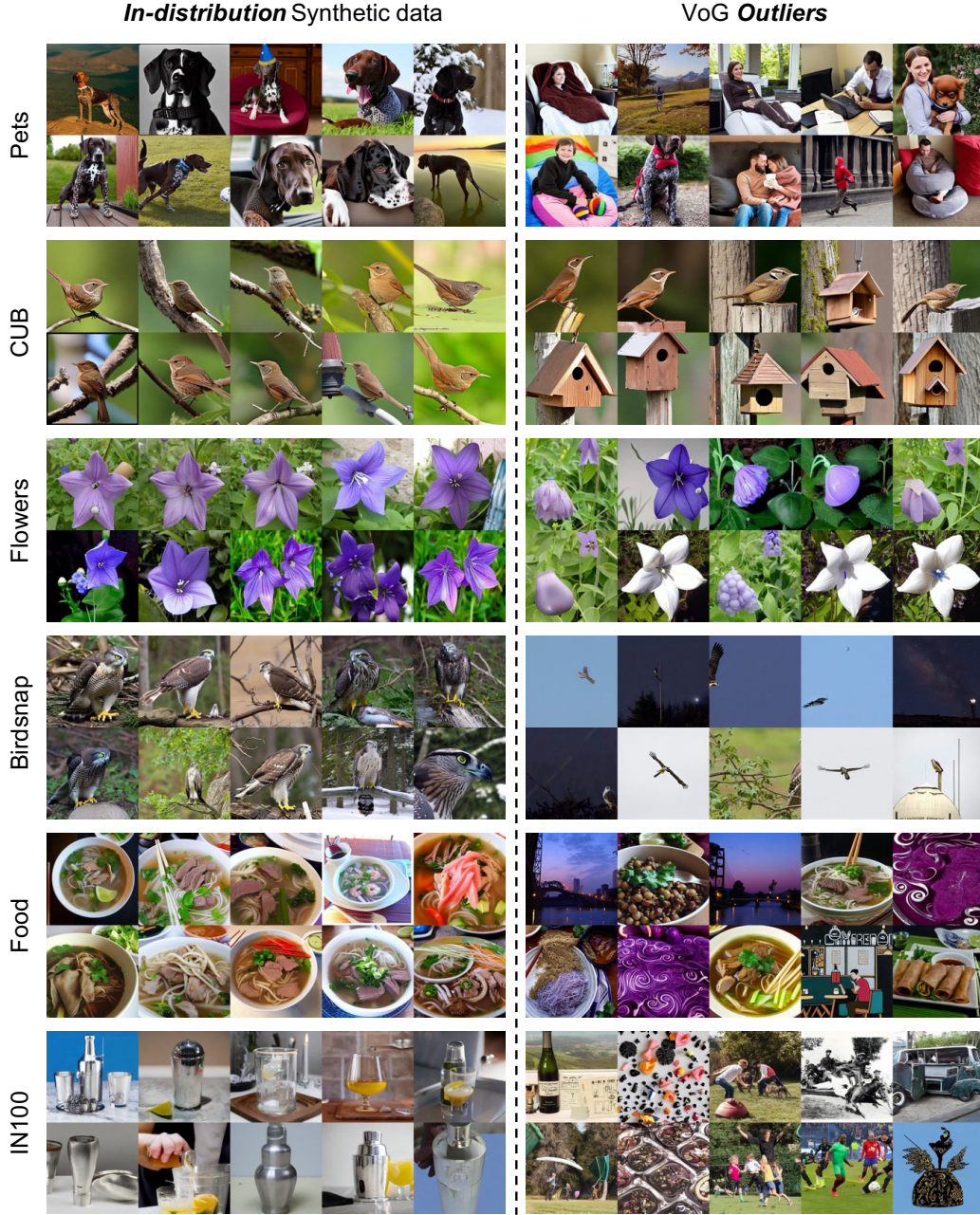


Figure 2: Comparison between in-distribution synthetic data and VoG outliers on all datasets.

image. Higher image strength results in noisier latent codes. We conducted experiments with Image Strength thresholds of 0.50, 0.75, and 0.90 (Table 2), using a default Marginal score Sampling Threshold and without VoG Filtering. Results indicate that Flowers prefer lower Image Strength for consistency with real data, while higher Image Strength benefits Pets, CUB, and Birdsnap datasets, highlighting the need for diversity in synthetic data.

Marginal score Sampling. In Table 3, we study the impact of different Marginal score Sampling thresholds without the VoG Filtering algorithm as well, where a specific portion of real data is sampled as marginal examples and the corresponding synthetic data is generated based on them. Although the optimal threshold for different datasets may vary, all threshold choices surpass the Real-Fake settings and show stability in performance.

Table 1: **Mean and standard deviation of Top-1 accuracy and worst-case disparity** with 3 different random seeds on ResNet-50 backbone.

Model	Pets	CUB	Flowers	Birdsnap
<i>Top-1 accuracy</i>				
Real-Fake [†] (Yuan et al., 2023)	94.0 ± 0.15	80.5 ± 2.25	89.2 ± 0.32	73.1 ± 0.31
Internet Explorer (15-NN similarity, (Li et al., 2023))	94.3 ± 0.25	83.2 ± 0.36	90.7 ± 0.44	73.7 ± 0.15
GenDataAgent (Ours)	94.6 ± 0.10	84.1 ± 0.17	91.5 ± 0.46	74.0 ± 0.57
<i>Worst-case disparity</i>				
Real-Fake [†] (Yuan et al., 2023)	0.48 ± 0.00	0.00 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
Internet Explorer (15-NN similarity, (Li et al., 2023))	0.51 ± 0.02	0.17 ± 0.07	0.54 ± 0.03	0.00 ± 0.00
GenDataAgent (Ours)	0.55 ± 0.02	0.25 ± 0.00	0.56 ± 0.00	0.00 ± 0.00

Table 2: Top-1 accuracy / Worst-case disparity of different Image Strength thresholds.

Image Strength	Pets	CUB	Flowers	Birdsnap	Food	IN100
0.50	94.0 / 0.44	81.0 / 0.25	89.9 / 0.50	71.4 / 0.00	87.4 / 0.63	88.6 / 0.40
0.75	94.0 / 0.44	82.2 / 0.25	88.4 / 0.40	73.6 / 0.00	87.4 / 0.63	88.7 / 0.40
0.90	94.4 / 0.56	83.6 / 0.25	88.8 / 0.40	73.5 / 0.00	87.4 / 0.61	89.1 / 0.40

Table 3: Top-1 accuracy / Worst-case disparity of different Marginal score Sampling thresholds.

Marginal score Threshold	Pets	CUB	Flowers	Birdsnap	Food	IN100
0 (only real)	93.6 / 0.40	83.1 / 0.13	87.4 / 0.40	73.0 / 0.00	86.8 / 0.63	87.4 / 0.20
1/5 training set	94.4 / 0.56	83.6 / 0.25	90.0 / 0.40	73.6 / 0.00	87.4 / 0.63	89.1 / 0.40
1/4 training set	94.2 / 0.48	83.9 / 0.25	89.9 / 0.50	73.5 / 0.00	87.5 / 0.63	89.6 / 0.40
1/2 training set	94.6 / 0.56	83.6 / 0.25	90.1 / 0.50	73.4 / 0.00	87.8 / 0.64	89.9 / 0.40
Full training set (Real-Fake)	94.2 / 0.48	83.1 / 0.00	89.0 / 0.50	73.0 / 0.00	87.4 / 0.61	88.6 / 0.40

Table 4: Top-1 accuracy / Worst-case disparity of different VoG Filtering thresholds.

VoG Filtering Threshold	Pets	CUB	Flowers	Birdsnap	Food	IN100
0% (Real-Fake)	94.2 / 0.48	83.1 / 0.00	89.0 / 0.50	73.0 / 0.00	87.4 / 0.61	88.6 / 0.40
25%	94.5 / 0.48	83.7 / 0.25	90.6 / 0.50	73.9 / 0.00	87.8 / 0.64	90.1 / 0.40
50%	94.5 / 0.48	83.9 / 0.25	90.1 / 0.50	74.2 / 0.00	87.6 / 0.61	89.6 / 0.40
75%	94.7 / 0.56	83.3 / 0.25	91.0 / 0.56	74.5 / 0.00	87.4 / 0.61	89.6 / 0.40
100% (only-real)	93.6 / 0.40	83.1 / 0.13	87.4 / 0.40	73.0 / 0.00	86.8 / 0.63	87.4 / 0.20

VoG Filtering. As shown in Table 4, different VoG Filtering ratios also show consistent improvement based on the Real-Fake setting, suggesting the removal of outliers is necessary with synthetic data augmentation. The preference for a relatively high filtering ratio suggests that the quality of synthetic data still has a gap compared to the real one, which can be a direction of future work.

E DATASETS AND TRAINING DETAILS

E.1 DATASET DETAILS

We adopt the general ImageNet-100 (IN100) dataset (Tian et al., 2020) and 5 popular fine-grained datasets: Oxford-IIT Pets (Parkhi et al., 2012), Flowers-102 (Nilsback & Zisserman, 2008), Birdsnap (Berg et al., 2014), CUB-200-2011 (Wah et al., 2011), and Food-101 (Bossard et al., 2014) for experiments. The dataset statistics are shown in Table 5.

E.2 TRAINING DETAILS

We use the same Stable Diffusion v1.5 (Rombach et al., 2022) as Real-Fake (Yuan et al., 2023). When adapting the stable diffusion to the target distribution, we use the full format "A photo of [Classname],

Table 5: Dataset statistics.

	Pets	CUB	Flowers	Birdsnap	Food	IN100
Domain	Pet Breed	Fine-grain Birds	Fine-grain Flowers	Fine-grain Birds	Fine-grain Foods	Natural Image
#Training Data	3,680	5,994	2,040	47,386	75,750	126,689
#Test Data	3,669	5,794	6,149	2,443	25,250	5,000
No. Classes	37	200	102	500	101	100

which is [Raw Image Caption]" as depicted in Section 3.3. Following Yuan et al. (2023), we adapt the stable diffusion with Low-Rank Adaptation (LoRA) with the same hyperparameters, as well as the same set of negative prompts "distorted, unrealistic, blurry, out of frame" for generations. For synthetic data augmentation, we use prompt guidance 7.5 for all datasets. For the only synthetic data setting, we use prompt guidance 2.0 for all datasets as suggested by Sarıyıldız et al. (2023). The hyperparameter for training the downstream classification model is listed in Table 6.

Table 6: Training hyperparameters of downstream classification model.

	Pets	CUB	Flowers	Birdsnap	Food	IN100
On-the-fly Iterations	20	20	20	20	20	20
Train Res \rightarrow Test Res	224 \rightarrow 224	448 \rightarrow 448	224 \rightarrow 224	224 \rightarrow 224	224 \rightarrow 224	224 \rightarrow 224
Training Epochs	200	200	200	200	200	200
Batch size	128 \times 8	64 \times 8	128 \times 8	128 \times 8	128 \times 8	128 \times 8
Optimizer	SGD	SGD	SGD	SGD	SGD	SGD
LR	0.1	0.2	0.1	0.1	0.1	0.1
LR decay	multistep	multistep	multistep	multistep	multistep	multistep
decay rate	0.2	0.2	0.2	0.2	0.2	0.2
decay epochs	50/100/150	50/100/150	50/100/150	50/100/150	50/100/150	50/100/150
Weight decay	5e-4	5e-4	5e-4	5e-4	5e-4	5e-4
Warmup epochs	-	-	-	-	-	-
Label smoothing	-	-	-	-	-	-
Dropout	x	x	x	x	x	x
CE loss \rightarrow BCE loss	x	x	x	x	x	x
Mixed precision	✓	✓	✓	✓	✓	✓

F MORE EXPERIMENTAL RESULTS

Accuracy under common corruptions and perturbations. To evaluate robustness, we introduced common corruptions such as Gaussian Blur and Speckle Noise, following the methodology outlined in Hendrycks & Dietterich (2019). All methods were trained on clean images and tested on corrupted images to assess their robustness. As shown in Table 7, our method consistently outperforms Real-Fake across all datasets and corruptions.

Table 7: Top-1 accuracy evaluated on Clean / Gaussian Blur / Speckle Noise images.

Top-1 Acc	Pets	CUB	Flowers	Birdsnap	Food	IN100
Real-Fake	94.2 / 93.0 / 90.4	83.1 / 77.3 / 75.4	89.0 / 88.0 / 79.2	73.0 / 72.4 / 71.7	87.4 / 86.6 / 86.1	88.6 / 88.1 / 87.2
GenDataAgent	94.7 / 93.6 / 90.9	83.9 / 79.1 / 76.0	91.0 / 90.3 / 82.4	74.5 / 73.8 / 72.4	87.8 / 87.1 / 86.7	90.1 / 89.7 / 89.0

Train and validation accuracy gap. We present the train accuracy, validation accuracy, and accuracy gap after convergence for all datasets in Table 8. Notably, the train-validation accuracy gap is reduced on GenDataAgent, which can be seen as a form of mitigation for overfitting. The reason might be the synthetic data covers some cases that the real data ignore (the diversity introduced by LLaMA caption perturbation), thus enhancing the model’s robustness and generalization capabilities.

G LIMITATIONS

Representation Ability of Generative Models. Our method augments real data with synthetic data generated by stable diffusion. Consequently, the classification model’s representation ability is

Table 8: **Train accuracy / validation accuracy / accuracy gap over all datasets.**

Train / Val / Gap	Pets	CUB	Flowers	Birdsnap	Food	IN100
Real-Fake	99.7 / 94.2 / 5.5	93.6 / 83.1 / 10.5	99.1 / 89.0 / 10.1	90.2 / 73.0 / 17.2	96.6 / 87.4 / 9.2	95.4 / 88.6 / 6.8
GenDataAgent	97.8 / 94.7 / 3.1	92.2 / 83.9 / 8.3	98.3 / 91.0 / 7.3	89.1 / 74.5 / 14.6	96.2 / 87.8 / 8.4	94.8 / 90.1 / 4.7

significantly influenced by the generative model’s capabilities, including different models and their versions.

Efficiency of Generating High-quality Synthetic Data. As shown in the main paper, generating synthetic data might be the bottleneck of synthetic data augmentation. Fast generative models with high-quality output are needed to improve efficiency. Despite techniques like those in [Bolya & Hoffman \(2023\)](#), generation time remains costly and quality can degrade.

Downstream Applications. Currently, synthetic data primarily augments classification tasks. We aim to expand GenDataAgent’s applicability to more downstream tasks, but this may require specific design adjustments for stable diffusion.

Adaptive Thresholds. The thresholds in GenDataAgent are manually set. Future work could explore more automated methods for setting these thresholds.

H BROADER IMPACTS

Most previous works ([Sarıyıldız et al., 2023](#); [Yuan et al., 2023](#)) studying synthetic data primarily focus on the quality of generative images, aiming to closely match the distribution of synthetic data with the target dataset. However, these studies often overlook the interaction between synthetic data and downstream classification. Our work bridges this gap by introducing Marginal score sampling and on-the-fly strategies, emphasizing the creation of **focused**, **diverse**, and **in-distribution** synthetic data.

REFERENCES

- Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2011–2018, 2014.
- Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4598–4602, 2023.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014.
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
- Alexander Cong Li, Ellis Langham Brown, Alexei A Efros, and Deepak Pathak. Internet explorer: Targeted representation learning on the open web. In *International Conference on Machine Learning*, pp. 19385–19406. PMLR, 2023.
- Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.
- 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*, pp. 10684–10695, 2022.
- Mert Bülent Saryıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8011–8021, 2023.
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. *arXiv preprint arXiv:2310.10402*, 2023.