LoFT: LoRA-Fused Training Dataset Generation with Few-shot Guidance

Jae Myung Kim[1,2,4] Stephan Alaniz[2,3,4] Cordelia Schmid[5] Zeynep Akata[2,3,4]

[1]University of Tübingen [2]Helmholtz Munich [3]Technical University of Munich
[4]Munich Center for Machine Learning [5]Inria, Ecole normale supérieure, CNRS, PSL Research University

Abstract

*Despite recent advances in text-to-image generation, using synthetically generated data seldom brings a significant boost in performance for supervised learning. Oftentimes, synthetic datasets do not faithfully recreate the data distribution of real data, i.e., they lack the fidelity or diversity needed for effective downstream model training. While previous work has employed few-shot guidance to address this issue, existing methods still fail to capture and generate features unique to specific real images. In this paper, we introduce a novel dataset generation framework named LoFT, **Lo**RA-**F**used **T**raining-data Generation with Few-shot Guidance. Our method fine-tunes LoRA weights on individual real images and fuses them at inference time, producing synthetic images that combine the features of real images for improved diversity and fidelity of generated data. We evaluate the synthetic data produced by LoFT on 10 datasets, using 8 to 64 real images per class as guidance and scaling up to 1000 images per class. Our experiments show that training on LoFT-generated data consistently outperforms other synthetic dataset methods, significantly increasing accuracy as the dataset size increases. Additionally, our analysis demonstrates that LoFT generates datasets with high fidelity and sufficient diversity, which contribute to the performance improvement. The code is available at https://github.com/ExplainableML/LoFT.*

1. Introduction

Synthetic data offers a cost-effective alternative to the labor-intensive process of real data collection. One promising downstream application of diffusion-based text-to-image generative models [5, 23, 35, 38, 39, 43] is to augment real datasets with synthetic images [13, 48] or training models on entirely synthetic data [19, 44, 58]. While these methods show potential, models trained solely on synthetic data often underperform compared to those trained on real data [14]. This is largely due to distributional misalignment between synthetic and real data, as well as a lack of fine-grained detail in the generated images [14, 25].

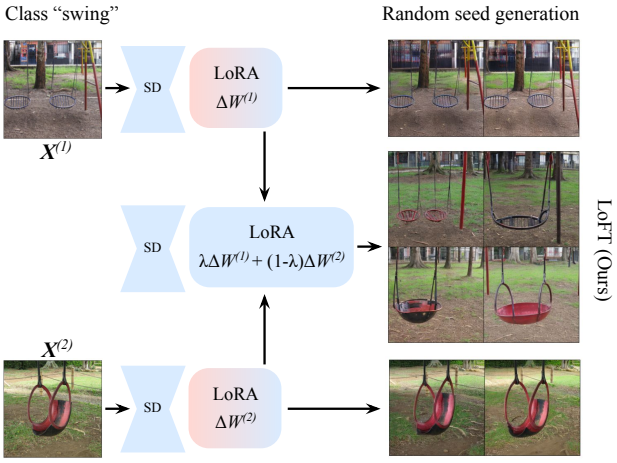


Figure 1. LoFT: Given a few real images per class, we first adapt a diffusion model to each image using LoRA. Next, two LoRA weights corresponding to images of the same class are randomly selected and fused to generate new images. The generated synthetic images above show diverse colors and compositions while maintaining the swing object.

To tackle distribution shift issue, recent works suggests to guide the dataset generation with a few real data samples [10, 19, 25]. In this few-shot setting, we assume to have access to a few real images for every class for an image classification task. Yet, the issue of misalignment remains a challenge for certain classes or downstream datasets. For instance, da Costa et al. [10] use partially noised real image as conditional input to the diffusion model and generate synthetic data using prompts from a captioning model. However, these images often deviate from the real data distribution, making them less relevant to the task at hand. Kim et al. [25] propose DataDream, which fine-tunes the diffusion model with few-shot data to learn the data distribution, but we find that it struggles to generate in-distribution images for all classes consistently. This is because DataDream finetunes on all available images from the same class, which

makes it challenging to retain high-fidelity details of individual images (e.g., less frequently visible parts of a class, such as the back of a car), focusing instead on commonly shared features.

We introduce LoFT, **Lo**RA-**F**used **T**raining-data Generation with Few-shot Guidance to generate high-fidelity, in-distribution synthetic images using few-shot real images. Instead of fine-tuning a diffusion model on all image of a class jointly, we train separate sets of Low-Rank Adaptation (LoRA) parameters on individual images, i.e., the diffusion model learns to overfit to a single image, generating it exclusively. At inference time, we then fuse together the LoRA weights of any two real images from the same class, to generate synthetic images that share the characteristics of both images. As shown in Figure 1, given two images of swings, the individual LoRA weights lead to generations similar to the real image (top and bottom), while the fused LoRA weights create images inheriting features from both source images while still maintaining the identity of a swing. There are two advantages of our LoFT method. First, learning separate LoRA weights for each individual image eases the diffusion model adaptation as the finetuning can retain on every detail of the real image. This instance-level adaptation ensures better alignment between the distribution of the few-shot real images and the synthetic images generated by the LoRA-tuned diffusion model, resulting in high fidelity. Second, by fusing the LoRA weights from different images of the same class, we maintain the diversity of the generated synthetic images.

Our key contributions are: (1) introducing LoFT, a few-shot guided synthetic dataset generation method that generates high-fidelity, in-distribution synthetic datasets by training LoRA adapters per image and fusing them when generating synthetic images; (2) providing a comprehensive comparison of four synthetic dataset generation methods on ten downstream datasets, demonstrating superior performance in fine-tuning CLIP when trained on data from LoFT; and (3) analyzing synthetic data generation methods based on fidelity and diversity, showing that LoFT achieves high fidelity with sufficient diversity, leading to improved performance when using its synthetic dataset.

2. Related Work

Diffusion-based text-to-image (T2I) models have enabled the creation of highly realistic synthetic images [5, 32, 35, 38, 39, 43]. These models operate by gradually denoising Gaussian noise, conditioned on textual prompts. Promising downstream applications of T2I generative models include generating training data for classification [2, 13, 15, 19, 25, 27, 44, 48, 49, 56, 58, 59, 62, 63, 65], handling long-tail distributions [21, 47, 55], data distribution shifts [3, 12], semi-supervised learning [57], representation learning [51], object detection [29], vision-language pre-training [17, 45, 52], and image generation [1, 4]. For image classification, a lot of work has focused on generating synthetic images zero-shot, which typically involves generating data from text prompts that include the class names from the downstream task [19, 44, 48, 58, 62]. However, this approach often leads to generated images that lack a faithful representation of the target object, resulting in a mismatch between synthetic and real images, which hinders performance gains [25, 52]. To mitigate these issues, there has been growing interest in few-shot learning, where limited real data is used alongside synthetic data. Techniques such as initializing the generation from a partially noised real image [10, 19] or fine-tuning the diffusion model with few-shot data [25] have been employed to align the synthetic data more closely with real-world distributions. Our method differs from other few-shot guided methods by using LoRA fusion to combine the features of multiple real images.

Controllable text-to-image diffusion models have enabled personalization in image generation [16, 36, 40, 41, 46, 53]. Fine-tuning a diffusion model with LoRA [24] and fusing them in the image generation phase has been demonstrated to be an effective technique for image morphing [60] and model customization [11], where LoRA weights are fused to achieve customized outputs. In contrast, our method leverages LoRA fusion for synthetic dataset generation and demonstrates its effectiveness in training classification models. While Zhou et al. [65] proposed fusing learned tokens from Textual Inversion [16] for synthetic dataset generation, we show that our LoFT outperforms these previous fusion methods.

To understand and improve the impact of synthetic training data, Fan et al. [14] measure the fidelity and diversity of synthetic datasets. Fidelity can be improved through methods like CLIP filtering [13, 19, 29] and incorporating additional class information [44], while diversity is affected by the guidance scale [14, 44], adding attributes to prompts [13, 44, 48], or using large language models to generate more varied prompts [17, 19]. Additionally, Fan et al. [14] have explored scaling laws in synthetic training data, demonstrating that synthetic datasets do not exhibit the same scaling benefits as real data in supervised tasks. In this work, we examine few-shot guided dataset generation methods in terms of both fidelity and diversity, and further investigate how scaling up synthetic datasets to sizes of up to one million affects the performance of these methods.

3. LoRA-Fused Training Dataset Generation

In this section, we begin by describing baseline methods for synthetic dataset generation in the zero-shot and few-shot scenarios (§3.1). We then introduce our proposed method, LoFT, in §3.2.

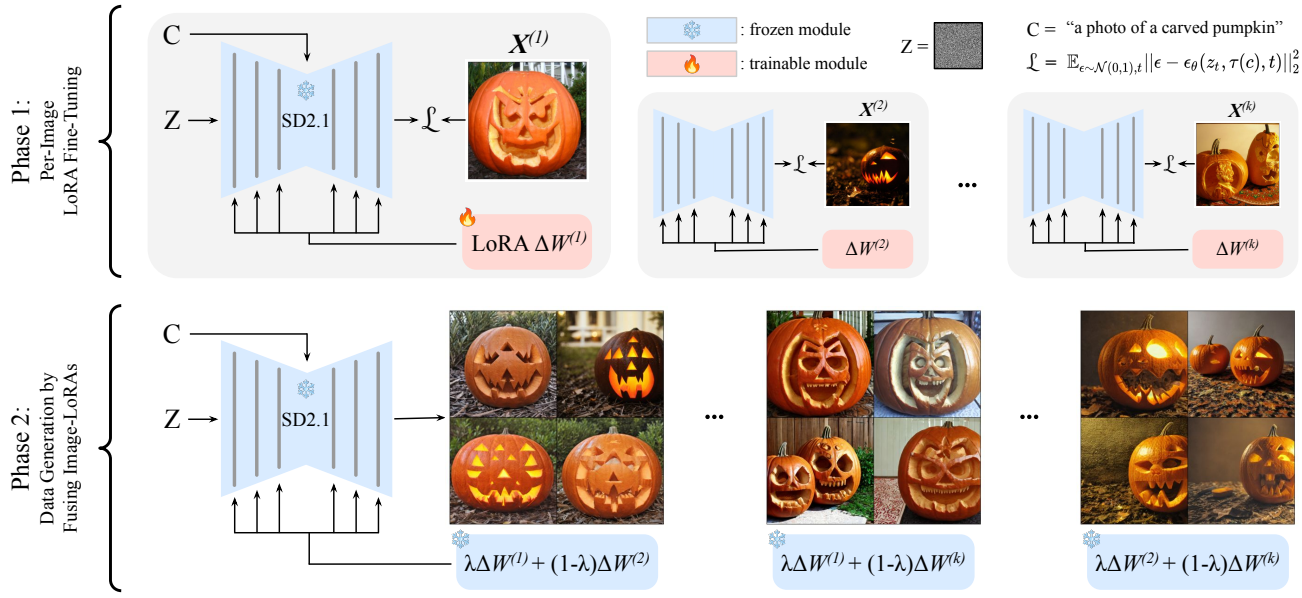


Figure 2. LoFT pipeline. In the first phase, given a few real images per class, we adapt a diffusion model to each image using LoRA. In the second phase, two LoRA weights corresponding to images of the same class are randomly selected and fused to generate new synthetic images. These generated images are then compiled to form a dataset for training the classification model.

3.1. Synthetic dataset generation

Stable diffusion: text-to-image generation. The Stable Diffusion [39] model learns a conditional probability distribution $p(x|c)$ given a data point $(x, c) \in \mathcal{D}$ where x is an image and c is its caption. The model learns a reverse process of gradually denoising Gaussian noise in the latent space. Concretely, the diffusion and reverse processes work in a latent space, which is defined through a pre-trained image encoder f that encodes the image x to a latent z, i.e. $z = f(x)$, and the corresponding decoder g where $x = g(z)$. Given a time step $t \in \{0, \cdots, T\}$, z_t denotes noisy latent state after t steps of small Gaussian noise addition from $z_0 = z$ where z_T is Gaussian noise. The latent diffusion models' objective is to minimize the following loss:

$$\min_{\theta} \mathbb{E}_{(x,c)\sim\mathcal{D},\,\epsilon\sim\mathcal{N}(0,1),\,t} \left[\| \epsilon - \epsilon_\theta(z_t, \tau(c), t) \|_2^2 \right], \quad (1)$$

where $\tau(\cdot)$ is a text encoder. Intuitively, the loss enables the model to learn to denoise the latent z_t. During inference, we start with noise z_T and iteratively denoise it through T steps of the latent diffusion model, obtaining z_0. This latent is decoded by the pre-trained decoder g to generate a final image $x' = g(z_0)$.

Zero-shot image generation. New images can be generated by conditioning the model on a template prompt [19, 44], such as "`a photo of a` $\{l\}$", where l represents a class name. As a result, the synthetic dataset

$$\mathcal{D}^{\text{synth}} = \{(x_i, y_i)\}_{i=1}^{sL}$$

contains s generated images for each of the L classes where every image x_i is automatically annotated by the class label $y_i \in \{1, 2, \cdots, L\}$ derived from its textual prompt. To improve diversity in these generated images, lowering the guidance scale has been shown to be effective, as it encourages more output variety, therefore improving classification performance [14, 44]. We refer to this method as **ClassPrompt**.

Few-shot guided dataset generation. While zero-shot text-to-image methods can generate a large amount of distinct images, they often struggle to produce the classification object of interest or capture fine-grained details of a class [25]. To address this, few-shot guided approaches have been developed, where we assume access to a few real images for each class. In the k-shot setting, we denote

$$\mathcal{D}^{\text{fs}} = \{(x_i, y_i)\}_{i=1}^{kC}$$

as the few-shot dataset, where x_i is an image, y_i is the label of the image, k is the number of available real images per class, and C is the number of classes. Few images per class can already provide rich visual information to better inform the data generation process beyond textual labels alone, while not requiring an extensive effort to collect.

For instance, DataDream [25] fine-tunes LoRA weights applied on the diffusion model using few-shot real images. In our experiments, we use DataDream with LoRA weights trained for each class as a representative of the few-shot guided image generation approach based on fine-tuning, and refer to this as **DataDream**.

While lowering the guidance scale increases variety in zero-shot image generation, using a template text prompt still limits the generation of a diverse dataset. To further increase diversity, Yu et al. [58] have leveraged large language models (LLMs) to enrich prompts with additional context or attributes related to the class name. Additionally, Dunlap et al. [13] and Fan et al. [14] leverage real images by applying a captioning model to create detailed captions from these images, which are then used as prompts for generation. In our experiments, we include a baseline for few-shot guided data generation through captioning. Specifically, we caption the few-shot images for each class with PaliGemma [6], a multimodal large language model. We generate one caption per real image, i.e., k captions per class in the k-shot setting. We then use these captions as prompts for synthetic image generation. We refer to this method as **CaptionPrompt**.

3.2. LoFT method

While DataDream has shown promising results for few-shot guided dataset generation, we find that it struggles to generate in-distribution images for some classes, limiting its impact on classification performance. This issue arises when there is high diversity in the few-shot images such that the fine-tuned diffusion models do not faithfully represent fine-grained details that may occur only in one of the images, leading to underfitting.

To overcome these challenges, we propose LoFT, **Lo**RA-**F**used **T**raining-data Generation with Few-shot Guidance for generating better in-distribution synthetic images. As shown in Figure 2, LoFT fine-tunes the pre-trained diffusion model with one set of LoRA weights for *every real image* x_i from the few-shot dataset $\mathcal{D}^{\text{fs}}$ independently. Specifically, for every attention layer of the diffusion model U-net, we add LoRA [24] parameters to the linear weight matrices

$$h_{\text{out}} = W h_{\text{in}} + \Delta W^{(i)} h_{\text{in}} \tag{2}$$

where h is the activation of a linear layer, $W \in \mathbb{R}^{d_1 \times d_2}$ is the original weight matrix, and $\Delta W^{(i)}$ is the low rank adaptation matrix which is optimized. The parameterization

$$\Delta W^{(i)} = B^{(i)} A^{(i)}$$

with $B \in \mathbb{R}^{d_1 \times r}$ and $A \in \mathbb{R}^{r \times d_2}$ allows for efficient fine-tuning because the low rank $r \ll min(d_1, d_2)$ reduces the number of tunable parameters significantly. The parameter efficiency and modularity of LoRA lead to a low storage costs and flexible inference-time manipulation (fusion).

Fine-tuning LoRA weights on a single image. As presented in the grey box in Figure 2, we fine-tune a separate set of LoRA weights $\Delta W^{(i)}$ for each data point (x_i, y_i) with the diffusion model objective while keeping the original parameters fixed:

$$\min_{\Delta W^{(i)}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t} \left[|| \epsilon - \epsilon_{\theta, \Delta W^{(i)}}(z_t, \tau(C(y_i)), t) ||_2^2 \right], \tag{3}$$

where z_t corresponds to the noised x_i at step t and $C(y_i)$ is the template prompt `"a photo of a {`l_i`}"` with l_i being the class name of y_i. By learning LoRA weights for every few-shot image individually, LoFT overfits the diffusion model to a single sample, learning to reproduce all of its details even if such features were not in the training data of the original diffusion model. At inference time, every generated image will closely resemble the original real image, increasing fidelity, i.e., replicating fine-grained details.

Fusing LoRA weights. To increase the generation diversity from single-image LoRAs, we propose to interpolate their weights. We fuse two randomly selected LoRA weights corresponding to real images from the same class with

$$h_{\text{out}} = W h_{\text{in}} + \lambda \Delta W^{(i)} h_{\text{in}} + (1 - \lambda) \Delta W^{(j)} h_{\text{in}} \tag{4}$$

where $\lambda \in [0, 1]$ and $\{(i, j) | y_i = y_j\}$.

This fusing strategy combines features from the different instances, effectively interpolating between real images in the weight space of the diffusion model, and improving the diversity of the generated images, as shown in the second phase in Figure 2. While $\lambda = 0$ or $\lambda = 1$ are reproducing the real data, choosing $\lambda = 0.5$ best interpolates images to produce new in-distribution samples of both high fidelity and diversity.

4. Experiments

We use Stable Diffusion [39] version 2.1 as the generative model for all synthetic dataset generation methods. For the ClassPrompt approach, we utilize the template prompt, `"a photo of a {`l`}"`. In the CaptionPrompt method, the prompt `"Caption the image:"`[1] is used to caption each input image with PaliGemma [6]. Once a list of captions is generated, each caption is appended to the template prompt, forming prompts such as `"a photo of a {`l`}, {caption}."` These prompts are then used as conditional input to the diffusion model to create synthetic images. For DataDream, We adopt the hyperparameter configuration Kim et al. [25] except that we exclude LoRA on text encoders for training since we found this to perform better. A guidance scale of 2.0 is used for all methods when generating synthetic images.

[1] Sourced from the original code base.

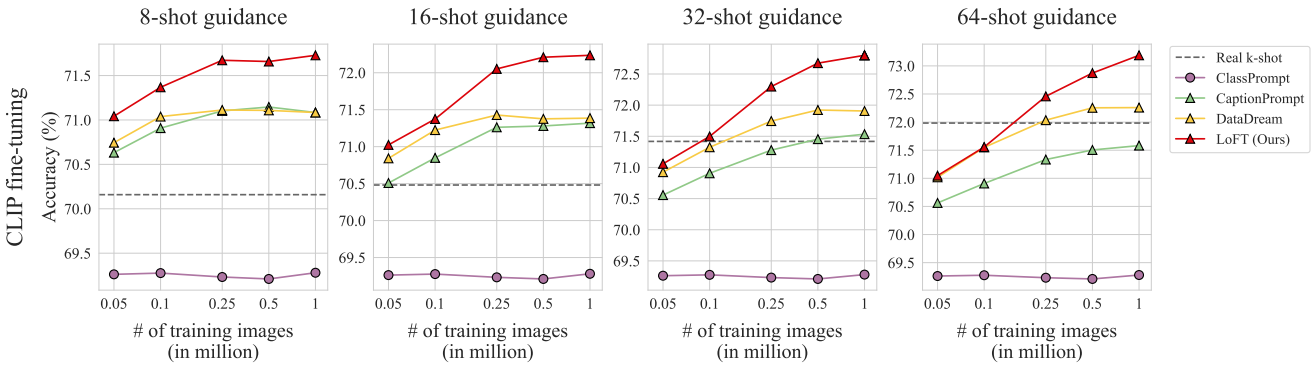


Figure 3. Classification accuracy on ImageNet when fine-tuning CLIP on synthetic data generated from different methods at different scales. We report few-shot guidance on 8, 16, 32, and 64 images per class and a baseline of training CLIP only on k-shot real data. LoFT consistently outperforms other methods and real k-shot result with small amount of synthetic data.

For our LoFT method, we employ AdamW [30] as the optimizer, a learning rate of 1e-3 with a cosine annealing scheduler, and a LoRA rank $r = 2$ for all trained LoRA adapters, which proved sufficient for adapting to single images. When generating images, LoFT fuses LoRA weights by randomly selecting two LoRA adaptations and fusing them with equal weights $\lambda = 0.5$. Different fusion strategies of LoRA weights are studied in §4.5.

4.1. Synthetic training data on ImageNet

To evaluate different dataset generation methods, we train an image classification model on each synthetic dataset. The target classification task is ImageNet [42], which contains 1,000 classes. For each generation method, we produce 50, 100, 250, 500, and 1,000 images per class, corresponding to dataset sizes of 0.05M, 0.1M, 0.25M, 0.5M, and 1M. We use these datasets to fine-tuning a pre-trained CLIP [37] model and evaluate it on the validation set of ImageNet. Our goal is to investigate whether synthetic training data can provide additional useful information to improve the performance of a pre-trained model that already has some knowledge of the downstream task. Following the work of da Costa et al. [10] and Kim et al. [25], we use the pre-trained CLIP ViT-B/16 model [37] as the base model, and fine-tune a LoRA (rank 16) applied to both the vision encoder and text encoder with the synthetic training data for ImageNet. We also conducted training ResNet50 [18] from scratch, which is shown in Appendix B. For synthetic data generation methods leveraging few-shot real images, we conduct experiments in 8-, 16-, 32- and 64-shot settings to examine performance under different guidance levels. The CLIP fine-tuning results are shown in Figure 3.

ClassPrompt does not scale. ClassPrompt dataset generation (purple line) improves over the baseline CLIP performance of 66.6%, and fluctuates between 69% and 70% as the dataset size increases. It indicates that the ClassPrompt method fails to improve as more images are generated. Consequently, synthetic images generated by the ClassPrompt method provide minimal additional information to the pre-trained CLIP model and showing no ability to scale.

Few-shot guided methods outperform ClassPrompt. Across all dataset sizes and k-shot settings, all few-shot guided methods (△ markers in Figure 3) outperform ClassPrompt (○ markers). This is because few-shot guided methods generate higher-quality images with better diversity (for CaptionPrompt) or higher fidelity (for DataDream and LoFT). We provide a detailed analysis in §4.3 and a qualitative examples in §4.4. This indicates that the inclusion of real-image guidance in the synthetic data generation process significantly improves the quality of the synthetic training dataset for training the downstream model.

Few-shot guided methods outperform real k-shot. The dashed line in each plot represents the performance of a model trained solely on k real images per class. We observe that even with a small number of synthetic images, models trained on a few-shot guided synthetic dataset can easily outperform models trained with real k-shot data. For example, the accuracy of the 16-shot real data is 70.48%. By generating 50 synthetic images per class (resulting in dataset size 0.05M), the accuracy of the model trained on LoFT dataset can reach 71.02%.

LoFT effectively scales across different k-shot settings. Unlike ClassPrompt, few-shot guided methods show consistent improvement in performance as the dataset size increases, with LoFT achieving the best performance across all k-shot settings. For instance, in the 16-shot setting, LoFT shows a 1.22% performance improvement, increasing from 71.02% to 72.24% as the dataset size grows from 0.05M to 1M training images, while CaptionPrompt shows a 0.81% improvement (70.51% → 71.32%). The difference becomes bigger in the 64-shot setting where LoFT shows a 2.14% improvement (71.05% → 73.19%) while CaptionPrompt shows a 1.02% improvement (70.56% → 71.58%).

Method	Cal	DTD	Eur	Air	Pet	Car	SUN	Food	Flo	Avg
CLIP zero-shot	93.0	44.4	47.6	24.7	89.2	65.2	62.6	86.1	71.4	64.9
ClassPrompt	$93.9_{\pm 0.2}$	$53.7_{\pm 0.5}$	$46.2_{\pm 0.7}$	$26.5_{\pm 0.1}$	$92.5_{\pm 0.1}$	$74.1_{\pm 0.5}$	$67.0_{\pm 0.0}$	$85.1_{\pm 0.0}$	$71.9_{\pm 0.5}$	$67.9_{\pm 0.1}$
CaptionPrompt	$95.4_{\pm 0.3}$	$63.9_{\pm 1.0}$	$46.6_{\pm 1.7}$	$26.6_{\pm 0.1}$	$92.9_{\pm 0.1}$	$74.4_{\pm 0.1}$	$73.9_{\pm 0.2}$	$85.4_{\pm 0.0}$	$72.3_{\pm 0.5}$	$70.2_{\pm 0.1}$
DataDream	$96.0_{\pm 0.4}$	$64.9_{\pm 0.2}$	$84.1_{\pm 3.4}$	$61.4_{\pm 0.9}$	$\mathbf{93.5}_{\pm 0.2}$	$\mathbf{90.5}_{\pm 0.4}$	$74.4_{\pm 0.2}$	$\mathbf{86.5}_{\pm 0.0}$	$\mathbf{98.0}_{\pm 0.2}$	$83.2_{\pm 0.4}$
LoFT (Ours)	$\mathbf{96.7}_{\pm 0.2}$	$\mathbf{70.5}_{\pm 0.1}$	$\mathbf{86.8}_{\pm 2.2}$	$\mathbf{66.1}_{\pm 1.5}$	$93.2_{\pm 0.1}$	$89.3_{\pm 0.5}$	$\mathbf{75.5}_{\pm 0.0}$	$86.0_{\pm 0.0}$	$\mathbf{98.0}_{\pm 0.2}$	$\mathbf{84.7}_{\pm 0.2}$

Table 1. Classification accuracy on 9 fine-grained benchmarks when fine-tuning CLIP on synthetic data with 16-shot guidance. 500 synthetic images are generated for each class. Datasets are Cal: Caltech 101, Eur: EuroSAT, Air: FGVC Aircraft, Flo: Flowers 102.

Method	Cal	DTD	Eur	Air	Pet	Car	SUN	Food	Flo	Avg
CLIP zero-shot [37]	93.0	44.4	47.6	24.7	89.2	65.2	62.6	86.1	71.4	64.9
CoOp [64]	$95.5_{\pm 0.1}$	$67.8_{\pm 2.2}$	$78.9_{\pm 0.4}$	$38.7_{\pm 0.7}$	$93.3_{\pm 0.3}$	$78.3_{\pm 0.7}$	$74.0_{\pm 0.2}$	$86.7_{\pm 0.6}$	$95.8_{\pm 0.1}$	$78.8_{\pm 0.6}$
TIP-Adapter [61]	$95.1_{\pm 0.1}$	$65.4_{\pm 1.2}$	$77.6_{\pm 1.0}$	$39.4_{\pm 0.3}$	$91.8_{\pm 0.3}$	$75.6_{\pm 0.5}$	$72.1_{\pm 0.2}$	$86.5_{\pm 0.1}$	$94.6_{\pm 0.1}$	$77.6_{\pm 0.2}$
TIP-Adapter-f [61]	$95.8_{\pm 0.1}$	$72.2_{\pm 0.3}$	$89.0_{\pm 0.4}$	$44.9_{\pm 0.4}$	$93.0_{\pm 0.2}$	$83.3_{\pm 0.5}$	$76.3_{\pm 0.2}$	$\mathbf{87.3}_{\pm 0.0}$	$96.8_{\pm 0.2}$	$82.1_{\pm 0.0}$
AMU-Tuning [50]	$97.1_{\pm 0.3}$	$70.0_{\pm 1.0}$	$90.4_{\pm 0.4}$	$47.7_{\pm 1.6}$	$92.8_{\pm 0.1}$	$78.5_{\pm 0.1}$	$72.6_{\pm 0.2}$	$85.7_{\pm 0.2}$	$95.4_{\pm 0.2}$	$81.1_{\pm 0.3}$
LoFT (Ours)	$\mathbf{97.3}_{\pm 0.1}$	$\mathbf{73.8}_{\pm 0.5}$	$\mathbf{93.1}_{\pm 0.9}$	$\mathbf{71.8}_{\pm 1.6}$	$\mathbf{94.3}_{\pm 0.4}$	$\mathbf{90.7}_{\pm 0.3}$	$\mathbf{77.3}_{\pm 0.0}$	$87.2_{\pm 0.1}$	$\mathbf{99.2}_{\pm 0.0}$	$\mathbf{87.2}_{\pm 0.3}$

Table 2. Comparison between the state-of-the-art few-shot learning methods on 9 fine-grained benchmarks. CLIP ViT-B/16 is used as a base model with a 16-shot setting. Baseline methods use real data, and LoFT use real data as well as synthetic data for the training set. Datasets are Cal: Caltech 101, Eur: EuroSAT, Air: FGVC Aircraft, Flo: Flowers 102.

4.2. Synthetic training data on fine-grained datasets

4.2.1. Comparison of synthetic data generation methods

To further examine the effectiveness of our LoFT method, we evaluate on 9 fine-grained benchmarks: Caltech101 [28], DTD [9], EuroSAT [20], FGVC Aircraft [31], Oxford Pets [34], Stanford Cars [26], SUN397 [54], Food101 [7], and Flowers102 [33]. For each dataset, we generate 500 synthetic images for each class and fine-tune the CLIP model. For few-shot methods, we use 16-shot.

The results are shown in Table 1. We observe that our LoFT method outperforms other methods on 6 out of 9 benchmarks, achieving the best average accuracy of 84.7%. For instance, LoFT performs significantly better than DataDream on the DTD dataset (70.5% vs. 64.9%) which consists of texture images. This may attribute from LoRA by DataDream having challenges in learning texture patterns with a batch of images, whereas optimizing a single image using LoFT leads to better convergence, thus generating more in-distribution images in the generation phase. Additionally, ClassPrompt and CaptionPrompt underperform on the Aircraft and Cars benchmarks, due to the limitations of diffusion models in distinguishing fine-grained classes based on their class names. We further study when scaling the number of synthetic images in Appendix C. It shows that for LoFT, the scaling curve meets a plateau at 500 images per class on DTD, while it keeps increasing over 5000 images per class on the Aircraft dataset.

4.2.2. Comparison with Few-shot learning methods

We compare our method with state-of-the-art methods in the few-shot learning literature. CoOp [64] optimizes the learnable token of textual input, TIP-Adapter [61] designs a cache model from the few-shot training set, and AMU-Tuning [50] balances the CLIP logit with MOCOv3 [8]. We reproduce the results using the official source code of each. While the few-shot learning methods optimize a pre-trained model using real data only, with LoFT method, we use both the few-shot real images as well as 500 synthetic images per class for the training set. CLIP ViT-B/16 is used as a base model, and we conduct experiments in the 16-shot setting.

The results are presented in Table 2. Our LoFT method outperforms baseline methods on 8 out of 9 benchmarks, achieving the best average accuracy of 87.2% vs. 82.1% of the next best method TIP-Adapter-f. We observe that there are significant performance gaps between LoFT and baseline methods on the Aircraft and Cars datasets, which consist of fine-grained classes. This indicates the advantages of using the synthetic data of LoFT in addition to the few-shot real images. Qualitative examples on these datasets can be found in Appendix D.

4.3. Per-class analysis on ImageNet

To identify how different factors in the synthetic dataset impact performance, following Fan et al. [14], we evaluate two metrics: recognizability and diversity. For our analysis, we randomly sample 50 images per ImageNet class for each synthetic dataset generation method.

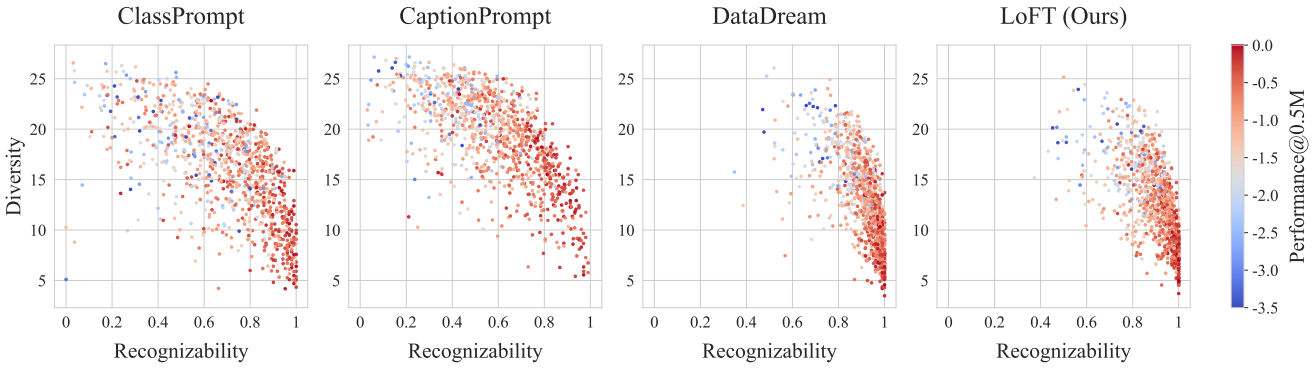
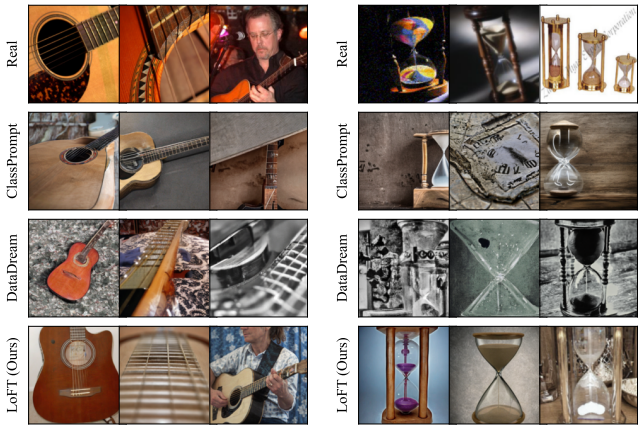


Figure 4. Per-class analysis on synthetic datasets generated from different methods. The color indicates a log-likelihood of the ImageNet validation dataset when CLIP is fine-tuned on the 0.5M-sized synthetic dataset in the 16-shot setting.



(a) Acoustic guitar (b) Hourglass

Figure 5. Qualitative examples for the classes Acoustic guitar and Hourglass from ImageNet. Our LoFT method generates diverse images, such as variations in zoom level, for acoustic guitar, and preserves an object of interest better for hourglass.

- Recognizability: To evaluate the fidelity of the generated images, we use a pre-trained ImageNet ViT-B/16 classifier (accuracy of 86.2%) to classify the generated images. The F1 score for each class serves as the metric.
- Diversity: For each class, we extract features from the same pre-trained ImageNet ViT-B/16 classifier and compute their standard deviation as a measure of diversity in the generated images.

Figure 4 presents the scatter plots for each method where each point summarizes one class. The color of each point indicates the log-likelihood of the corresponding class in the validation set of ImageNet, as predicted by the CLIP model fine-tuned on 0.5M synthetic images in the 16-shot setting.

Recognizability and diversity are inversely correlated. Across all methods, there is an inverse correlation between recognizability and diversity: as recognizability increases, diversity tends to decrease, and vice versa.

Few-shot guided methods exhibit higher recognizability, while other methods have higher diversity. DataDream and LoFT show higher recognizability compared to ClassPrompt and CaptionPrompt. This is because these few-shot guided methods are specifically trained to generate images similar to the real images. This alignment improves the realism and quality of the generated images, leading to higher recognizability. While LoFT incorporates multiple LoRA weights to increase diversity, it still shows less diversity than ClassPrompt and CaptionPrompt.

Distinct strengths of each method. CaptionPrompt obtains a good performance on classes with high diversity (i.e., in the range of 20-25). On the other hand, DataDream and LoFT demonstrate better performance with high recognizability (i.e., when it is greater than 0.8). This suggests that each method has its own strengths. It remains an open question for further exploration of methods that combine these strengths, achieving high diversity and recognizability.

4.4. Qualitative comparison

We present qualitative results in Figure 5 to gain insights into the diversity and quality of images from different methods. For the acoustic guitar class (Figure 5a), real images have high variety, including differences in zooming, color palettes, and the presence of humans. In contrast, ClassPrompt images lack this diversity, displaying limited color variation and similar representations of the guitar. DataDream demonstrates a better level of diversity, generating images with various colors and textures. However, some DataDream images exhibit artifacts, which can detract from their overall quality. In contrast, our LoFT method successfully balances diversity and image quality. This is reflected in the quantitative metrics, where LoFT achieves the highest scores for recognizability (0.94) and diversity (14.68), outperforming ClassPrompt (0.88, and 9.54).

In the hourglass category in Figure 5b, ClassPrompt sometimes produces image related to “hour” but not “hour-

Fusing representation	0.05M	0.1M	0.25M	0.5M
Caption embeddings	70.36	70.55	70.62	70.66
Image embeddings	69.88	69.95	69.99	70.02
Tokens from Textual Inversion	69.70	70.13	70.21	70.29
LoRA weights (= LoFT, ours)	**71.02**	**71.37**	**72.05**	**72.21**

Table 3. Comparison of methods on fusing different representations when fine-tuning CLIP. Experiments are done in the 16-shot setting on ImageNet.

glass", as seen in the second column. Some of the images by DataDream show barely recognizable shapes of an hourglass. In contrast, images from LoFT closely resemble the object of interest. This observation is consistent with the quantitative metrics where LoFT achieves a recognizability score of 1.0 while ClassPrompt and DataDream score 0.89 and 0.95, respectively.

4.5. Ablation study of LoFT

Our ablation study investigates how different methods of fusing representations impact the synthetic training dataset. We conduct two studies: one exploring the effect of various fusion techniques, and another examining the influence of the weight parameter λ in our LoRA-based fusion method.

4.5.1. Fusion on different representations

To explore the effectiveness of different fusion techniques on the resulting synthetic dataset, we compare three methods for fusing representations: 1) caption embedding fusion, which involves averaging the text embeddings of the PaliGemma captions from two images; 2) image embedding fusion, which directly embeds two images using an image encoder, and then passes the average image representation to Stable-Diffusion-2.1-unclip[2], a model for image-to-image generation; and 3) Textual Inversion [16] fusion, which optimizes input tokens for each image and then fuses the learned token embeddings from two images.

We generate up to 500 images per class on ImageNet using each method and fine-tune CLIP on the resulting datasets. Our results in the 16-shot setting are presented in Table 3, which shows that our LoRA-based fusion method outperforms the other techniques both when generating fewer samples (50K imgs, +0.66%) and with increasing number of generations (500K imgs, +1.55%). This suggests that our method is more effective at capturing the underlying structure of the data and generating high-quality images.

4.5.2. λ variation for LoRA fusion

We also examine the impact of λ on the generated images and downstream classification performance. In Appendix G we show examples of images generated with different values of λ illustrating that $\lambda = 0.5$ provides the best visual results, especially in terms of diversity.

In Table 4, we quantitatively demonstrate the importance of choosing an optimal value for λ. When λ is set to 0.5, our method achieves the best performance, with a significant increase in accuracy as the amount of training data increases (scaling from 29.03% to 45.41% with 0.05M to 0.5M samples). In contrast, setting λ to values closer to 0 or 1 results in lower performance, due to a lack of diversity in the generated images (only scaling from 22.42% to 30.85% with 0.05M to 0.5M samples for $\lambda = 1$).

Fusing LoRAs	0.05M	0.5M
$\lambda = 0.5$	29.03	**45.41**
$\lambda = 0.7$ (or 0.3)	25.60	39.18
$\lambda = 1$ (or 0)	22.42	30.85
$\lambda \sim \text{Beta}(2, 2)$	28.28	43.15
$\lambda \sim \text{Beta}(5, 5)$	28.56	44.91
$\lambda \sim \text{Beta}(10, 10)$	**29.40**	45.36
$[0.5, 0.25, 0.25]$	27.87	43.63
$[0.33, 0.33, 0.33]$	28.46	44.10
$[0.7, 0.15, 0.15]$	22.19	36.28

Table 4. Ablation study on LoRA fusion. We train ResNet50 from scratch in the 16-shot setting.

Second, we explore the impact of introducing randomness to the λ value, using a Beta distribution $\text{Beta}(\alpha, \alpha)$, where α controls the concentration of λ around the value of 0.5. Larger values of α lead to a distribution that is more concentrated around 0.5, while smaller values of α allow for a broader spread across the [0,1] interval. When α is small, the performance decreases, i.e., accuracy of 43.15% ($\alpha = 2$) vs. 45.36% ($\alpha = 10$) with 0.5M data samples. This suggests that having a more concentrated distribution around $\lambda = 0.5$ is beneficial for performance.

Finally, we also evaluate the performance of fusing three LoRA weights during dataset generation. Our results, presented in the last three rows of Table 4, show that none of these methods outperform the two-LoRA fusion with $\lambda = 0.5$. In fact, using more than two LoRAs introduces artifacts into the generated images, which can deteriorate their recognizability.

5. Conclusion

In this paper, we introduced LoFT, **Lo**RA-**F**used **T**raining-data Generation with Few-shot Guidance. LoFT fine-tunes LoRA weights on individual real images, ensuring high fidelity when generating synthetic images, and then fuses them to achieve diversity. Our experiments demonstrate that LoFT consistently outperforms other methods when fine-tuning a pre-trained CLIP model. This is because the synthetic images generated by LoFT complement the prior knowledge contained in CLIP by accurately capturing and fusing the features of each class. Additionally, we showed that LoFT performs better as the number of few-shot samples increases when training from scratch. Our analysis of the synthetic datasets showed that LoFT achieves a balance of high fidelity with reasonable diversity, while methods like ClassPrompt and CaptionPrompt focus more on generating diverse images at the cost of fidelity.

[2]https://huggingface.co/stabilityai/stable-diffusion-2-1-unclip

Acknowledgements

Jae Myung Kim thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD programs for support. This work was partially funded by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation, which we thank for their generous support. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *ICLR*, 2023. 2

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *TMLR*, 2023. 2

[3] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023. 2

[4] Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *ICLR*, 2024. 2

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1, 2

[6] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 4

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 6

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 6

[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 6

[10] Victor G. Turrisi da Costa, Nicola Dall'Asen, Yiming Wang, Nicu Sebe, and Elisa Ricci. Diversified in-domain synthesis with efficient fine-tuning for few-shot classification, 2023. 1, 2, 5

[11] Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. *arXiv preprint arXiv:2406.09413*, 2024. 2

[12] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *NeurIPS*, 2023. 2

[13] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation, 2023. 1, 2, 4

[14] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *CVPR*, 2024. 1, 2, 3, 4, 6

[15] Yunxiang Fu, Chaoqi Chen, Yu Qiao, and Yizhou Yu. Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*, 2024. 2

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2022. 2, 8

[17] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 12

[19] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 1, 2, 3

[20] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *AEROS*, 2019. 6

[21] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023. 2

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 12

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[24] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 4

[25] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. *ECCV*, 2024. 1, 2, 3, 4, 5

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, pages 554–561, 2013. 6

[27] Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image captions are natural prompts for text-to-image models. *arXiv preprint arXiv:2307.08526*, 2023. 2

[28] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 6

[29] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *CVPR*, 2023. 2

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 5

[31] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 6

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 6

[34] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 6

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2023. 1, 2

[36] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *NeurIPS*, 2023. 2

[37] 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 *ICML*, 2021. 5, 6

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *CVPR*, 2024. 2

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1, 2

[44] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, 2023. 1, 2, 3

[45] Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. Synth2: Boosting visual-language models with synthetic captions and image embeddings. *arXiv preprint arXiv:2403.07750*, 2024. 2

[46] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2

[47] Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023. 2

[48] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion, 2023. 1, 2

[49] Zhaorui Tan, Xi Yang, and Kaizhu Huang. Semantic-aware data augmentation for text-to-image synthesis. In *AAAI*, 2024. 2

[50] Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Amu-tuning: Effective logit bias for clip-based few-shot learning. In *CVPR*, 2024. 6

[51] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS*, 2023. 2

[52] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15887–15898, 2024. 2

[53] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *ICCV*, 2023. 2

[54] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 6

[55] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, Nayeong Kim, Suha Kwak, and Tae-Hyun Oh. Synaug: Exploiting synthetic data for data imbalance problems. *arXiv preprint arXiv:2308.00994*, 2023. 2

[56] Teresa Yeo, Andrei Atanov, Harold Benoit, Aleksandr Alekseev, Ruchira Ray, Pooya Esmaeil Akhoondi, and Amir Zamir. Controlled training data generation with diffusion models. *arXiv preprint arXiv:2403.15309*, 2024. 2

[57] Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. Diffusion models and semi-supervised learners benefit mutually with few labels. *NeurIPS*, 2023. 2

[58] Zhuoran Yu, Chenchen Zhu, Sean Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *arXiv preprint arXiv:2312.02253*, 2023. 1, 2, 4

[59] Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. In *ICLR*, 2024. 2

[60] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *CVPR*, 2024. 2

[61] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022. 6

[62] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, 2023. 2

[63] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. *NeurIPS*, 2023. 2

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 6

[65] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023. 2

Supplementary Material for
LoFT: LoRA-Fused Training Dataset Generation with Few-shot Guidance

A. Implementation details of classifier training

When fine-tuning the CLIP model, we freeze the CLIP parameters and update the LoRA weights applied to them, with a lora rank of 16. We use a batch size of 256 and a learning rate of 1e-6 with cosine annealing for the learning rate schedule. The weight decay is set to 1e-4. As the dataset size increases, we adjust the number of iterations to be increased while decreasing the number of epochs. For dataset sizes of 0.05M, 0.1M, 0.25M, 0.5M, and 1M, the number of epochs is set to 90, 80, 70, 60 and 50, respectively. The warm-up period is set to 10% of the total epochs, resulting in 9, 8, 7, 6, and 5 warm-up epochs for each dataset size.

When training the ResNet50 from scratch, we use a batch size of 2048 and a learning rate of 0.2 with cosine annealing for the learning rate schedule. The weight decay is set to 1e-4. As the dataset size increases, we adjust the number of iterations to be increased while decreasing the number of epochs. For dataset sizes of 0.05M, 0.1M, 0.25M, 0.5M, and 1M, the number of epochs is set to 300, 250, 200, 150, and 100, respectively. The warm-up period is set to 10% of the total epochs, resulting in 30, 25, 20, 15, and 10 warm-up epochs for each dataset size.

B. Training an image classifier from scratch

We further evaluate the dataset generation methods by training ResNet50 [18] model from scratch and evaluating it on the validation dataset of ImageNet. The results are shown in Figure 6.

Performance improves with dataset size. Contrary to CLIP fine-tuning, training from scratch with data from the ClassPrompt method improves with increasing dataset size. Similarly, the performance of all few-shot guided generation methods also improves consistently with data scale for all k-shot settings.

The best method depends on the k-shot setting. The best method differs depending on the number of k-shot real images used for guidance. In the case of smaller k-shot settings (8-shot), CaptionPrompt outperforms both DataDream and LoFT. However, as the number of shots increases, DataDream and LoFT start to outperform CaptionPrompt, with LoFT consistently performing better than DataDream. For instance, LoFT achieves 58.70% at 1M scale in 64-shot while DataDream and CaptionPrompt achieve 56.00% and 52.48%, respectively.

Conclusions differ between fine-tuning and training from scratch. While the best method for training from scratch depends on the number of k-shot real images used for guidance, LoFT outperforms all baseline methods consistently across all k-shot settings when fine-tuning the CLIP model. Synthetic images by LoFT complement the prior knowledge contained in CLIP because it generates images that accurately represent the features of each class. In contrast, CaptionPrompt introduces greater diversity but since CLIP has already been pre-trained on a broad range of diverse images, it does not provide as much complementary value, limiting its effectiveness for fine-tuning.

C. Scaling up to 5000 images per class on fine-grained datasets

To study the scaling ability of synthetic data size on fine-grained dataset, we conduct experiments by generating up to 5000 images per class for the Aircraft and DTD datasets. For the few-shot synthetic data generation methods, we use 16-shot real images as guidance. The results from fine-tuning CLIP are shown in Figure 7. ClassPrompt and CaptionPrompt reach a plateau from the beginning, indicating that increasing the number of synthetic images does not improve the classification performance. On DTD, LoFT reaches a plateau at 500 images per class, while on the Aircraft, performance continues to improve up to 5000 images per class.

D. Qualitative results on fine-grained datasets

We show qualitative results of our LoFT method on Aircraft and Cars datasets. For the DHC-8-100 class on the Aircraft dataset in Figure 15a, LoFT generate a propeller attached to the wing, which resembles real images. Moreover, for the Model B200 class in Figure 15b, LoFT generates the shape of the class similar to real images, such as the head shape and the tail shape. Similarly in Figure 15c and Figure 15d, LoFT generate images of the class "Jeep Wrangler SUV 2012" and "Bugatti Veyron 16.4 Coupe 2009" that resemble the shape and fine-grained details of real images, respectively.

E. Additional per-class analysis

Correlation between diversity and alignment. In addition to the recognizability and diversity metrics introduced in §4.3, we introduce one additional metric, **alignment**, to measure how closely the distribution of synthetic data aligns with that of real data. To quantify this, we calculate the Fréchet Inception Distance (FID) [22] score for each class, where a lower score indicates closer alignment between the synthetic and real data distributions.

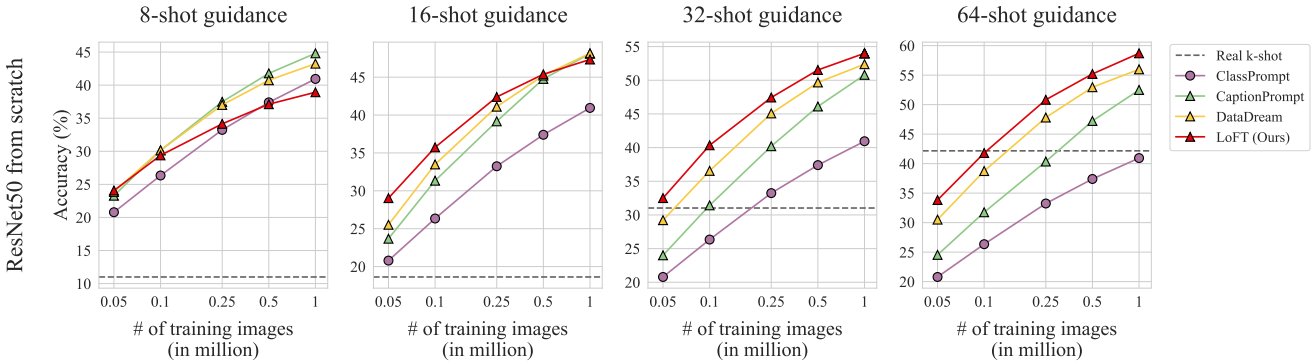


Figure 6. Classification accuracy on ImageNet when training ResNet50 from scratch on synthetic data generated from different methods at different scales. We report few-shot guidance on 8, 16, 32, and 64 images per class and a baseline of training CLIP only on k-shot real data. LoFT benefits from a larger number of real images as guidance.

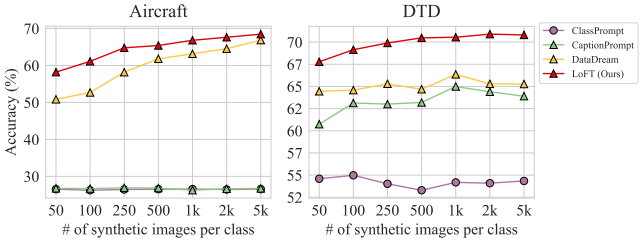


Figure 7. Scaling the number of synthetic data on Aircraft and DTD datasets when fine-tuning CLIP.

As seen in Figure 11, we observe a positive correlation between alignment and diversity for all methods. This suggests that higher diversity in the generated images come at the expense of closely mimicking the real data distribution. Moreover, the overall alignment scores are smaller for LoFT and DataDream compared to ClassPrompt and CaptionPrompt, indicating that the images generated by LoFT and DataDream align more closely with the real data distribution.

F. Ratio of data points the two models disagree on the prediction

Since different methods show distinct strengths that contribute to performance gains, a natural question arises: do the classification models trained on each synthetic dataset exhibit different sets of corrected data points? To explore this, we use a ResNet50 model trained on the 0.5M-sized dataset from the ClassPrompt and 16-shot guided generation methods. We calculate the ratio of the number of data points in the validation ImageNet dataset that show inconsistent predictions relative to the total number of data points, i.e. where one model makes a correct prediction while the other model makes an incorrect one.

The results are shown in Figure 8. Even though the three few-shot guided methods (CaptionPrompt, DataDream, and LoFT) have comparable overall accuracy (around 45% accuracy, in Figure 6), the correction flip ratios between them are above 20%. This suggests that each synthetic dataset encourages the model to learn different features. Moreover, LoFT shows a higher flip ratio with CaptionPrompt (26.6%) compared to DataDream (20.4%). This aligns with our per-class analysis, where CaptionPrompt maintains performance by leveraging greater diversity in image distribution, while LoFT and DataDream have higher recognizability, focusing more on image fidelity.

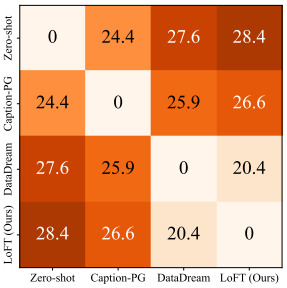


Figure 8. Ratio of data points the two models disagree on the prediction.

G. Qualitative comparison on ImageNet

We present additional qualitative results for 8 classes, i.e. Hourglass, Hard disk drive, Joystick, Weighing scale, Carved Pumpkin, Diaper, Swing, and iPod, in Figure 12 and Figure 13.

Taking the class Hourglass in Figure 12a as an example, real images show hourglasses with diverse frames and varying sand colors. The images generated by ClassPrompt show less color variation. While CaptionPrompt and DataDream generate more colorful images, some of them are not easily recognizable as hourglasses. In contrast, LoFT generates images that maintain both diversity in the frame and sand color while clearly representing the hourglass.

Taking the class Swing in Figure 13c as another example, real images show one or multiple swings, sometimes with a person riding them. Some of the generated images by ClassPrompt does not look like a swing, but rather resem-

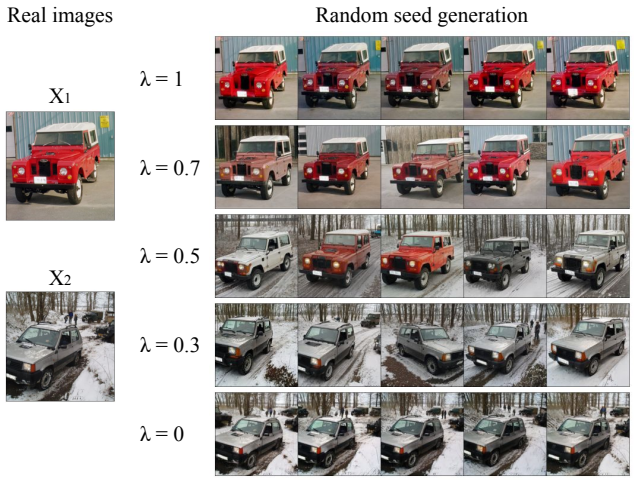


Figure 9. Ablation study of qualitative results on λ variation when fusing LoRAs. Given two images of Jeep class, $\lambda = 0.5$ merges features from both real images while maintaining diversity with random seed image generation. As λ approaches 0 or 1, the generated images become more similar to the original image and looses diversity.

ble a chair. For CaptionPrompt and DataDream, some of the generated images focus more on the human subject than the swing itself, making the swing less visible. In contrast, LoFT generates clear images of swings or multiple swings, with the object clearly identifiable.

H. Additional qualitative results varying λ

Figure 9 presents examples of images generated by our LoFT method with different λ values, alongside their corresponding real images. As we adjust the weight parameter λ for the LoRA fusion, we observe distinct trends in the generated outputs. When λ is set to either 0 or 1, the generated images closely resemble the original real images. However, this approach limits the diversity of outputs across different seeds. As λ approaches 0.5, we achieve an optimal balance that enhances the diversity of the generated images while preserving their quality. Each generated image effectively integrates features from the two original real images while resembling in-distribution data. This characteristic makes the synthetic training dataset produced by LoFT beneficial for classification tasks.

We present additional qualitative results in λ variation for 4 classes, i.e. Hourglass, Carved Pumpkin, Diaper, and Swing, in Figure 14.

Taking the class Hourglass in Figure 14a as an example, a real image x_1 shows a single hourglass with a wooden frame while another real image x_2 shows multiple hourglasses without a wooden frame. When $\lambda = 1$ or 0, the images generated by different random seeds closely resemble one of the real images. When $\lambda = 0.5$, the generated images show both diversity and high fidelity: some images have wooden frame while others do not, and some display multiple multiple hourglasses while others show only a single hourglass.

Taking the class Swing in Figure 14d as another example, a real image x_1 shows a baby riding a swing colored with yellow and blow while another real image x_2 shows only a yellow swing. When $\lambda = 1$ or 0, the images generated by different random seeds closely resemble one of the real images. When $\lambda = 0.5$, the generated images show both diversity and high fidelity: the color of the swing is different, and a baby is riding a swing in some of the images.

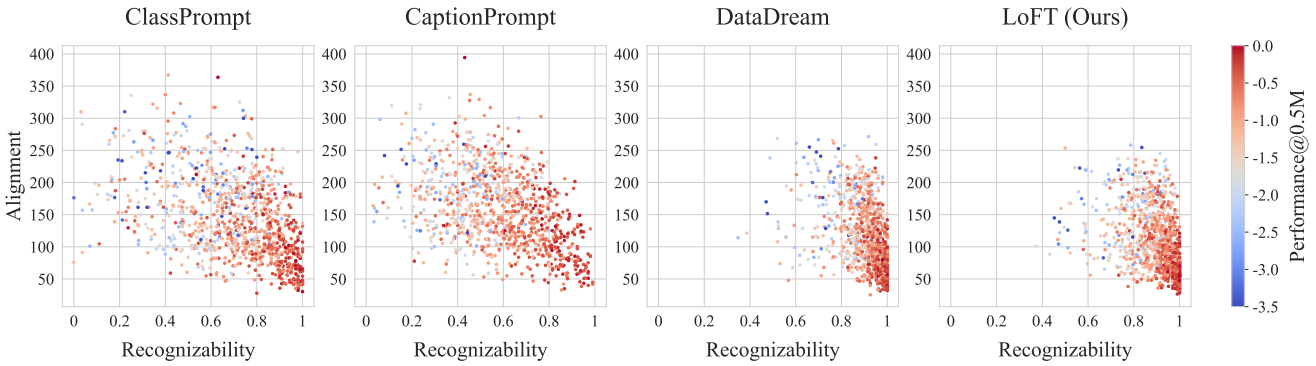


Figure 10. Per-class analysis of recognizability and alignment in synthetic datasets generated from different methods. The color indicates a log-likelihood of the ImageNet validation dataset when CLIP is fine-tuned on the 0.5M-sized synthetic dataset in the 16-shot setting.

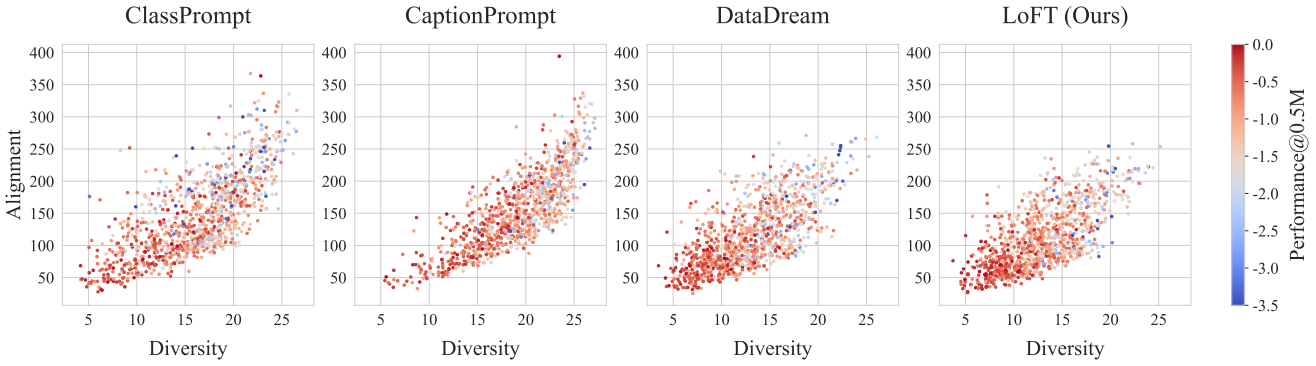
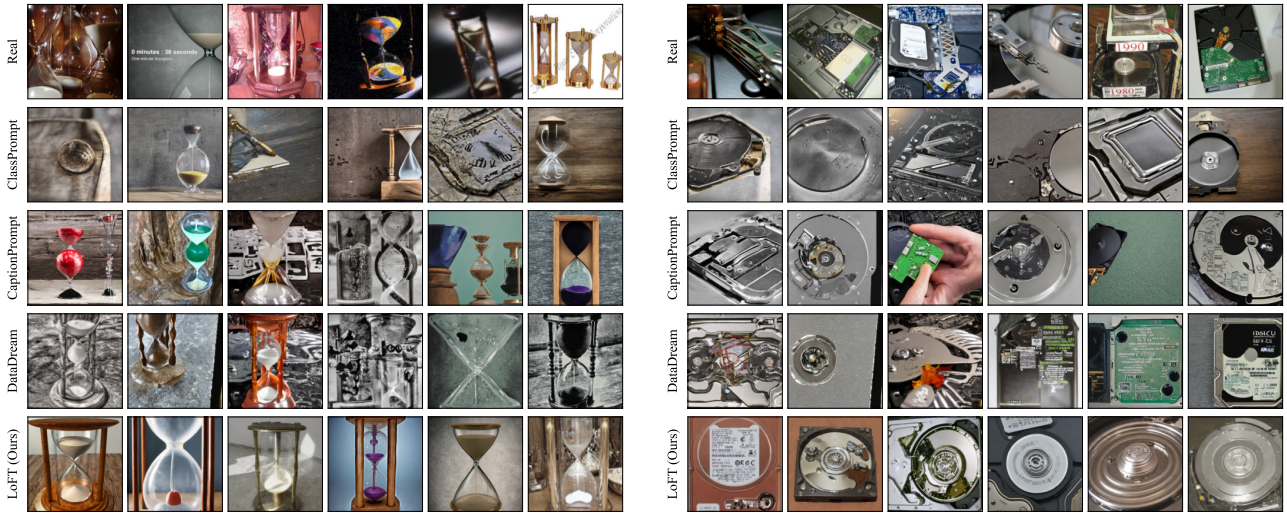


Figure 11. Per-class analysis of diversity and alignment in synthetic datasets generated from different methods. The color indicates a log-likelihood of the ImageNet validation dataset when CLIP is fine-tuned on the 0.5M-sized synthetic dataset in the 16-shot setting.



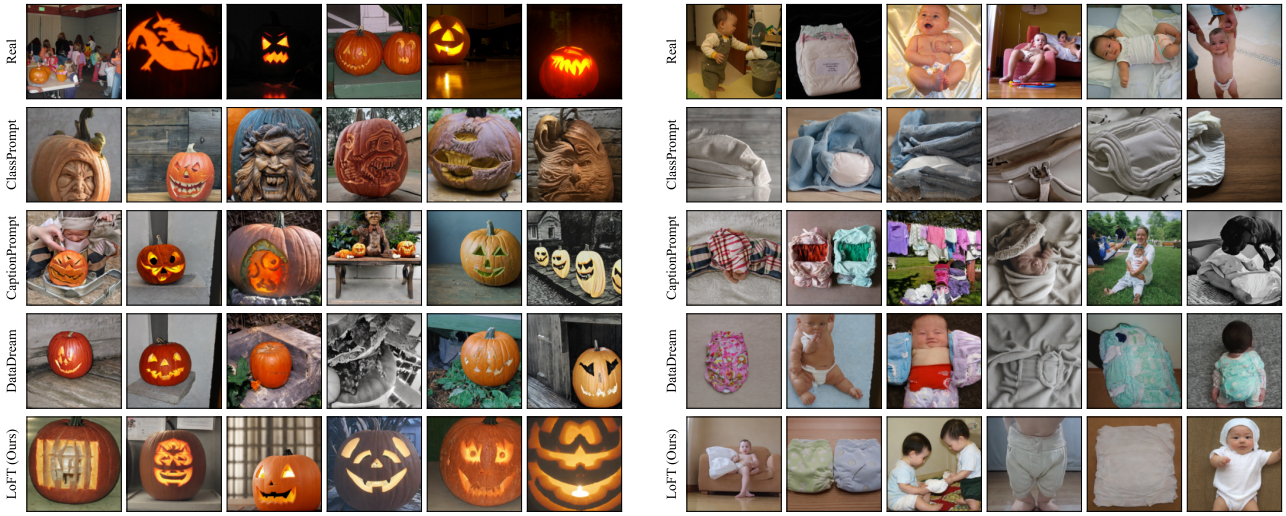
(a) Hourglass

(b) Hard disk drive

(c) Joystick

(d) Weighing scale

Figure 12. Qualitative examples for the classes Hourglass, Hard disk drive, Joystick, and Weighing scale.



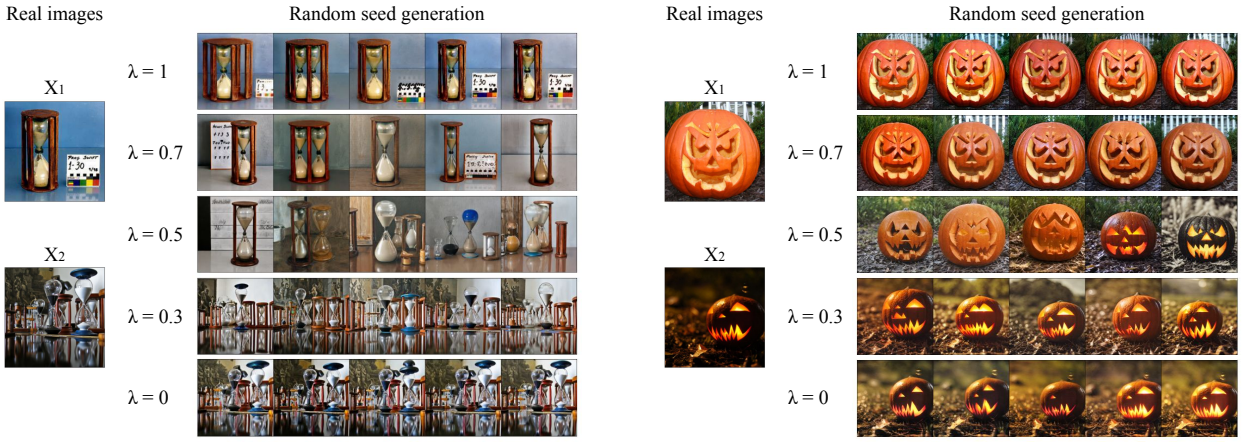
(a) Carved Pumpkin

(b) Diaper

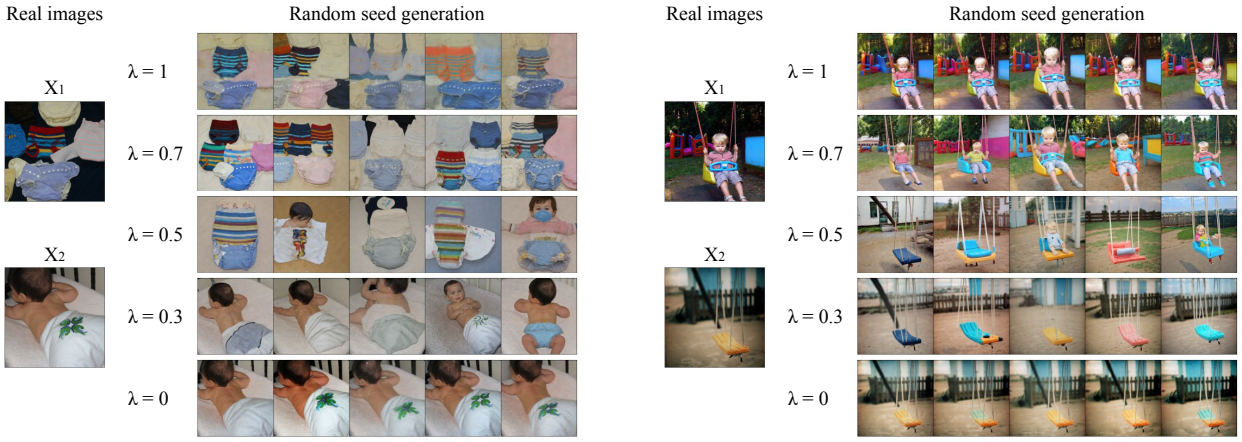
(c) Swing

(d) iPod

Figure 13. Qualitative examples for the classes Carved Pumpkin, Diaper, Swing, and iPod.



(a) Hourglass (b) Carved pumpkin



(c) Diaper (d) Swing

Figure 14. Ablation study of qualitative results on λ variation when fusing LoRAs.

(a) “DHC-8-100” class on Aircraft

(b) “Model B200” class on Aircraft

(c) “Jeep Wrangler SUV 2012” class on Cars

(d) “Bugatti Veyron 16.4 Coupe 2009” class on Cars

Figure 15. Qualitative results of our LoFT method on Aircraft and Cars datasets.