Generating Images of Rare Concepts Using Pre-trained Diffusion Models

Dvir Samuel[1,2*], **Rami Ben-Ari**[2], **Simon Raviv**[1], **Nir Darshan**[2], **Gal Chechik**[1,3]

[1]Bar-Ilan University, Ramat-Gan, Israel
[2]OriginAI, Tel-Aviv, Israel
[3]NVIDIA Research, Tel-Aviv, Israel

Abstract

Text-to-image diffusion models can synthesize high quality images, but they have various limitations. Here we highlight a common failure mode of these models, namely, generating uncommon concepts and structured concepts like hand palms. We show that their limitation is partly due to the long-tail nature of their training data: web-crawled data sets are strongly unbalanced, causing models to under-represent concepts from the tail of the distribution. We characterize the effect of unbalanced training data on text-to-image models and offer a remedy. We show that rare concepts can be correctly generated by carefully selecting suitable generation seeds in the noise space, using a small reference set of images, a technique that we call SeedSelect. SeedSelect does not require retraining or finetuning the diffusion model. We assess the faithfulness, quality and diversity of SeedSelect in creating rare objects and generating complex formations like hand images, and find it consistently achieves superior performance. We further show the advantage of SeedSelect in semantic data augmentation. Generating semantically appropriate images can successfully improve performance in few-shot recognition benchmarks, for classes from the head and from the tail of the training data of diffusion models.

1 Introduction

Diffusion models achieve unprecedented success in text-to-image generation. They map a noise vector sampled from a high-dimensional Gaussian, conditioned on a text prompt, to a corresponding image (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022; Balaji et al. 2022). While successful, several failure modes of current models have been identified. Common failures range from omitting objects listed in the prompt or confusing their attributes (Chefer et al. 2023; Rassin et al. 2023), through ignoring spatial relations (Lian et al. 2023) to generating deformed hands as illustrated in Figure 1.

One failure mode received less attention so far: some concepts and object classes consistently fail to be drawn correctly. Figure 1 illustrates these failures for two concepts: "pay phone" and "oxygen mask" in images generated with StableDiffusion (Rombach et al. 2022). These failures occur mostly with concepts that appear less frequently, but it is still not well understood what causes these failures, and if at all they can be corrected.

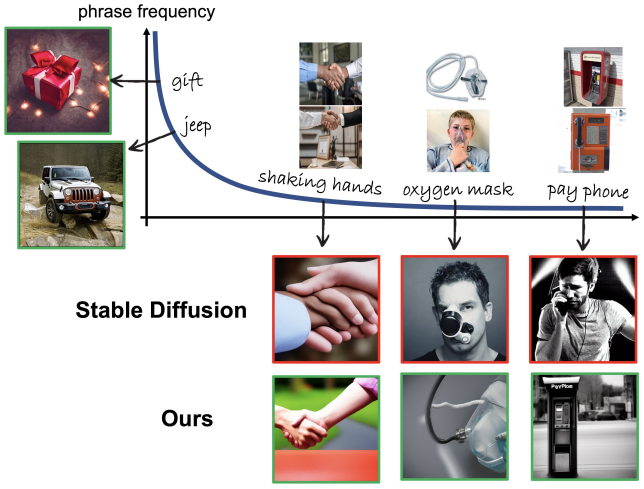

Figure 1: Generating rare concepts. Current diffusion models fail when conditioned on phrases or classes that are in the tail of their training distribution, like *pay phone*, or structurally complex classes like *shaking hands*. SeedSelect fixes that using just a handful of additional references images, without any fine-tuning.

Here we study a major failure mode of text-to-image diffusion models: generation of concepts that are under-represented in the training data. We first quantify this effect in a public model (Stable Diffusion) trained with public data (Schuhmann et al. 2022, Laion2B), and find that 25% of ImageNet concepts are poorly generated (Figure 2). The failing concepts are those that have fewer than 10K samples in the training data of the diffusion model. This observation is somewhat puzzling. Intuitively, 10k samples should be sufficient for learning the appearance of a concept, even a complex one.

Why do diffusion models fail to generate images from concepts with several thousand image samples? One possible answer raises from failures of deep models trained for *long-tail* recognition (Zhang et al. 2021). There, common concepts dominate the learned representation, washing out

*Correspondence to: Dvir Samuel <dvirsamuel@gmail.com>

the representation of rare concepts. If this type of "catastrophic forgetting" is the cause for the above failures in generative models, little can be done to improve the generation of rare concepts.

This paper explores a different answer. Our insight is that diffusion models may be sensitive to the initial random noise used as input in conjunction with their text prompts. When a diffusion model is trained for frequent concepts ("A dog"), its training covers a large fraction of the random input space. The model then learns to map any noise sample correctly to viable images. In contrast, for rare concepts, only a small fraction of that input space is observed during training. As a result, at generation time for a given prompt, the model may view many random inputs as out-of-distribution.

Based on this view, we show that, indeed, diffusion models can generate images from rare concepts, as long as the initial noisy image (the seed) is carefully chosen. To achieve this, we use a small set of reference images from the class. We identify areas in the noise space (seeds) that would be "in-distribution" for the diffusion model for a given prompt. More concretely, we do a gradient-based search in input space for regions that generate images that are similar, visually and semantically, to our few-shot reference set. We call our approach **SeedSelect**.

We evaluate the quality of images generated with SeedSelect in several ways. First, we evaluate the faithfulness of generation, namely, if generated images depict the correct class. This is done (a) using a classifier that was pre-trained to recognize each concept and (b) using human raters. SeedSelect consistently achieves better faithfulness than all competing approaches, for concepts that are in the tail of the Laion2B distribution, across three datasets (Imagnet, iNaturalist and CUB). SeedSelect also achieves better image quality as measured using FID. Then, we test the benefit of using generated images for semantic augmentation in an object recognition task. SeedSelect achieves state-of-the-art results for few-shot image recognition tasks on ImageNet, CUB, and iNaturalist. It generates valuable, diverse, and superior augmentations compared to previous methods. Finally, SeedSelect can be used to improve generation of challenging concepts, such as hand palms, where current diffusion models struggle.

Our paper makes the following contributions: (1) We characterize the failure of text-to-image diffusion models to generate images of rare concepts. (2) We introduce the learning setup of rare-concept generation using a reference set and a pre-trained text-to-image diffusion model. (3) We describe *SeedSelect*, a novel method to improve generation of uncommon and ill-formed concepts in diffusion models. It operates as per-class test-time optimization by finding a generation seed from just a few reference samples. (4) We propose an efficient bootstrapping technique to accelerate image generation with SeedSelect.

2 Related Work

Text-guided generation: Diffusion models provide unprecedented quality for text-to-image generation (Ramesh et al. 2022; Saharia et al. 2022; Balaji et al. 2022) but still struggle with rare fine-grained objects and compositions (Chefer et al. 2023; Liu et al. 2022). Techniques like pre-trained image classifiers (Dhariwal and Nichol 2021) and text-driven gradients (Ho and Salimans 2021; Nichol et al. 2022; Saharia et al. 2022) have been proposed for better aligning generated images with the given text prompt, but require pre-trained classifier which may not be available or extensive prompt engineering (Liu and Chilton 2022; Marcus, Davis, and Aaronson 2022; Wang et al. 2022). Other approaches (Avrahami et al. 2023; Feng et al. 2023; Chefer et al. 2023) generate more accurate images, and focus on aligning better the generated images to the prompt, not addressing the generation of rare objects. Our approach also improves alignment with the prompt, in the sense of forcing the model to generate a well-formed or correct image, particularly when the concept is rare.

Semantic augmentations for image recognition with pre-trained text-to-image models: Recently, (He et al. 2023; Azizi et al. 2023) showed that data augmentations obtained from images generated by pre-trained text-to-image models improve zero-shot and few-shot image classification. (He et al. 2023) achieves SOTA results by fine-tuning a CLIP classifier with real and synthetic images. Two strategies were introduced for generating images resembling few-shot reference images: (1) Real Guidance (RG) guides image generation using few-shot real samples, where these samples (with added noise) replace initial random noise to steer the diffusion process. (2) Real Filtering (RF) uses few-shot real sample features to filter similar synthetic images. However effective, we show that these strategies compromise image diversity and naturalness and aren't suitable for generating rare concepts. (Azizi et al. 2023) showed that large-scale text-to-image diffusion models, when fine-tuned, can produce class conditional models that enable classifiers trained on such generated data to excel in classification benchmarks. Despite their effectiveness, this approach demands substantial fine-tuning data, which might be lacking, especially for generating rare concepts. Our approach, on the other hand, demonstrates how to generate such rare concepts without finetuning the diffusion model.

Image generation personalized to an instance: Recently, (Gal et al. 2023a; Ruiz et al. 2023; Tewel et al. 2023) described how few reference samples can be used to train a model to generate images of a unique instance object. In principle, these methods can also be used for generating rare concepts and for few-shot semantic data-augmentation. However, they require long-time training for a single concept, and importantly, they do not learn a "class concept" but rather an "instance-specific concept" (or style), as in "this *specific* cat" and not "this *type* of cat". Accelerated versions like (Gal et al. 2023b) are limited to specific classes. Overall, these methods require substantial computational resources. Thus, when evaluating them as data augmentation methods, we only compare our approach with Textual Inversion (Gal et al. 2023a) on the CUB (Wah et al. 2011) dataset.

3 Motivating Analysis

We start by quantifying the relation between two quantities: the faithfulness of images generated for a given class by a

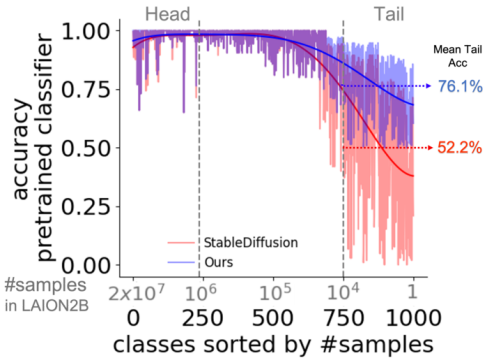

Figure 2: Per-class accuracy of a pre-trained classifier for images generated using stable diffusion. Shown are the 1000 classes of ImangeNet1k ordered by their number of occurrences in the LAION2B dataset.

common text-to-image model, and the number of samples from that class in the training set.

Foundation diffusion models are trained on massive datasets, collected "in the wild" from the web (Schuhmann et al. 2022). The distribution of concepts in web images is highly unbalanced, with some concepts appearing orders-of-magnitude more frequently than others. As a result, trained diffusion models are well-tuned to "head" concepts, but when asked to generate images from "tail" classes, the results are poor. Figure 1 illustrates how this imbalance is manifested in the LAION2B dataset (Schuhmann et al. 2022). We parsed all image captions and extracted all noun phrases in each caption (more details and a similar analysis of LAION400M are given in the supplementary. Data will be publicly released).

We quantified the relation between faithfulness and training imbalance with the following experiment. For every class in ImageNet (Deng et al. 2009), we used Stable Diffusion (Rombach et al. 2022, SD) to generate 100 images using the class label as the prompt. We then used a SoTA pre-trained classifier provided by Tu et al. (2022) to test if the generated images are from the correct class (see supplementary for details). That classifier was trained on balanced ImageNet data and has no preference for classes that appear at the head of the Laion distribution. Figure 2 depicts the resulting per-class accuracy for ImageNet classes sorted by their prevalence in the LAION2B dataset. Images generated for categories at the head of LAION2B distribution yield high accuracy, but accuracy drops significantly at the tail, particularly for classes in the last quartile (last 25% of classes). For those rare classes, about 50% of synthetic images generated by Stable Diffusion are correctly identified, indicating corrupted or incorrect concepts. In the figure, we also report the mean accuracy of classes from the last quartile (mean tail acc). Note that concepts from many tail classes were observed thousands of times in the Laion training data. This behavior strongly limits the usability of diffusion models to generate rare concepts.

Since the diffusion model was trained with thousands of samples from rare classes, a natural question arises: **Are these classes encoded in the model? and if so can they be revived and generated?** or were they washed out by the overwhelmingly many more samples from head classes?

Our working hypothesis: Deep diffusion models are trained given two inputs: a text prompt, and a noisy image, which in the extreme case is a noise tensor sampled from a high-dimensional Gaussian distribution. We propose that when trained with common (head) concepts, the model learns to map large parts of that Gaussian distribution into images of correct concepts. However, for rare (tail) concepts, the model can generate correct concepts only for limited areas of that distribution. If that is true, then if we can locate these areas of the distribution, we could still generate images of rare concepts. In this paper, we propose to discover these areas by optimizing over the seed in the noise-space, such that it improves semantic and appearance agreement with a small set of reference images of target rare concepts. Figure 2 shows that images generated by our approach achieve better faithfulness. In the subsequent sections, we will elaborate on the details of our method.

4 Notations and Definitions

We start with defining the problem of rare-concept generation with a reference set. Given a pre-trained text-to-image model (like StableDiffusion), a rare concept y to be generated, and k reference images $I^1, I^2, ...I^k$ of the concept y, the objective is to generate new and semantically-correct images of y.

While our approach can be directly applied to all diffusion models, in this work we use the open-sourced model of Stable Diffusion (SD) (Rombach et al. 2022). In the context of SD, a denoising diffusion probabilistic model (DDPM) is applied to the latent space of a variational auto-encoder. The process involves training an encoder $\mathcal{E}$ to map images to spatial latent codes z, and a decoder $\mathcal{D}$ to reconstruct images from these codes. The DDPM, informed by conditioning vectors (often derived from pre-trained CLIP text encoders), operates on the latent space and uses a network ε_θ to effectively remove noise ε from the latent code z using UNet architecture with self-attention and cross-attention layers. During inference, a latent z_T is sampled from a standard normal distribution and iteratively denoised with DDPM to yield a latent z_0, which is then decoded by $\mathcal{D}$ to generate the final image I^G. More details in Supp.

5 Our Approach: Seed Select

We now describe how we use *few reference images*, $I^1, \ldots, I^k$, to improve generation of images for a given prompt y. Typically, k can be set to 3-5 samples. Our goal is to find an initial noise tensor z_T^G that generates images that are consistent with the reference set as illustrated in Figure 3. We measure this consistency in two ways:

(1) Semantic consistency. Measures the semantic similarity between the generated image I^G obtained from a seed z_T^G and the reference images $I^1, \ldots, I^k$. Specifically, we use a pre-trained CLIP image encoder to encode the reference images into $v^1, \ldots v^k$, and compute their centroid (mean vector): $\mu_v = mean(v^1, \ldots, v^k)$. Similarly, we encode the

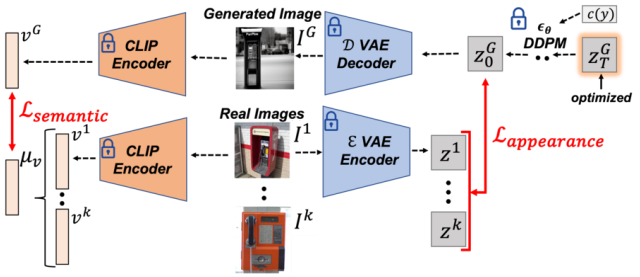

Figure 3: An overview of SeedSelect. An initial noise z_T^G is used to generate an image I^G. It is then tuned to minimize a semantic loss (using clip image encoder) and an appearance loss (using the diffusion VAE) based on its match to reference samples $I^1, \ldots, I^k$.

generated image and obtain v^G. The semantic loss is then:

$$\mathcal{L}_{Semantic} = dist_v(\mu_v, v^G), \tag{1}$$

where $dist_v$ is the euclidean distance between the centroid μ_v and a feature vector v^G. This loss makes sure that the semantic concept in the generated image corresponds to the concept presented in the reference images.

(2) Natural appearance consistency. Measures the similarity between the spatial latent z_0^G obtained from z_T^G during the denoising process and the encoded reference images. More specifically, we encode the reference images $I^1, \ldots, I^k$ using $\mathcal{E}$, the VAE encoder, to obtain $z^1, \ldots, z^k$. Then, we define the appearance loss to be:

$$\mathcal{L}_{Appearance} = \frac{1}{k}\sum_{i=1}^{k} dist_z(z^i, z_0^G), \tag{2}$$

where $dist_z$ is pixel-wise mean-squared error loss. Note that this loss mirrors the loss used during training of the diffusion model, namely the MSE between the latent tensor of the generated image and the latent representation of the provided real images. This loss mechanism ensures that the generated images maintain a natural appearance consistent with the provided images.

The overall loss is then:

$$\mathcal{L}_{Total} = \lambda\mathcal{L}_{Semantic} + (1-\lambda)\mathcal{L}_{Appearance} \quad . \tag{3}$$

Here, λ is a hyperparameter to control the tradeoff between semantic and appearance.

We only optimize z_T^G, the initial generation point, by backpropagating the loss through the denoising model, maximizing appearance and semantic consistency.

Implementation details: We use Stable Diffusion v2.1 with a guidance scale of 7.5 and 7 denoising steps using EulerDiscreteScheduler (Karras et al. 2022). See full implementation details in supplementary material.

Stopping criteria: We stop optimizing z_T^G when $\mathcal{L}_{Total}$ plateaus or its value increases for more than 3 iterations.

Inference (Image generation): Once an optimal z_T^G is found, generating an image is done by following the standard denoising process of the DDPM to obtain I^G. To generate multiple different images one can repeat the optimization by sampling a new z_T^G and optimize it using SeedSelect. See a faster method below.

5.1 Improving Speed and Quality

Stabilized optimization. The last few denoising steps z_t,z_{t-1}, .. z_0 for $t << T$, often generate high quality images. To speed up convergence, we compute the losses for all images in the last t steps $\mathcal{L}^t_{Semantic}$, and then aggregate them $\mathcal{L}_{Semantic} = \sum_{i=0}^{t} \mathcal{L}^i_{Semantic}$. In our experiments, we found $t = 2$ to be suitable to stabilize optimization.

Faster generation using bootstrap. Typically, finding an optimal z_T^G takes between 1-4 minutes on an NVIDIA A100 GPU. To quickly generate a large number of images, we operate as follows. First, execute the optimization procedure, with fewer iterations, to find an optimal z_T^G for the full set $\mathcal{I} = \{I_1, \ldots, I_k\}$. Then, use bootstrap (Efron 1992) to sample a subset $S \subset \mathcal{I}$ of reference images. Finally, find an optimal $z_T^{G_S}$ for the subset S, but start the optimization from z_T^G and generate the image I^{G_S}. We repeat this process for multiple subsets to obtain a diverse set of images. We find this bootstrap procedure, where we first learn a good initialization point and then generate images based on subsets, reduces the optimization duration for a single image from minutes to seconds.

Contrasting classes. When generating images from a set of classes C, we can further improve optimization convergence and image quality by using a supervised contrastive loss (Khosla et al. 2020). The loss operates in the *semantic space*; it pulls the semantic vector v^G closer to the centroid of its class μ_v^c, and pushes it away from centroids of other classes $\mu_v^{c'}$. The updated semantic loss is

$$\mathcal{L}_{Semantic} = -\log \frac{e^{-dist(\mu_v^c, v^G)}}{\sum_{c' \in C} e^{-dist(\mu_v^{c'}, v^G)}} \quad . \tag{4}$$

6 Experiments

To assess the quality of images generated for rare concepts, we analyzed several important aspects: faithfulness (whether the correct concept was generated), visual appeal (realism and naturalism of the image), diversity of generated images, and applicability to downstream applications.

To evaluate the faithfulness of generated images, we used SoTA pre-trained classifiers to determine whether the images belong to the correct class or not, supplemented by human evaluations. For assessing image realism, we used the FID score to quantify the distinction between real and generated samples.

For diversity, we used standard measures to find the Precision, Recall, Fidelity, and Diversity of generative models.

Finally, we assessed how SeedSelect can benefit two downstream applications: (1) for generating hand palm images, a challenging task since foundation diffusion models were published. and (2) as semantic data augmentations to enhance few-shot CLIP classification. An ablation study can be found in Supp.

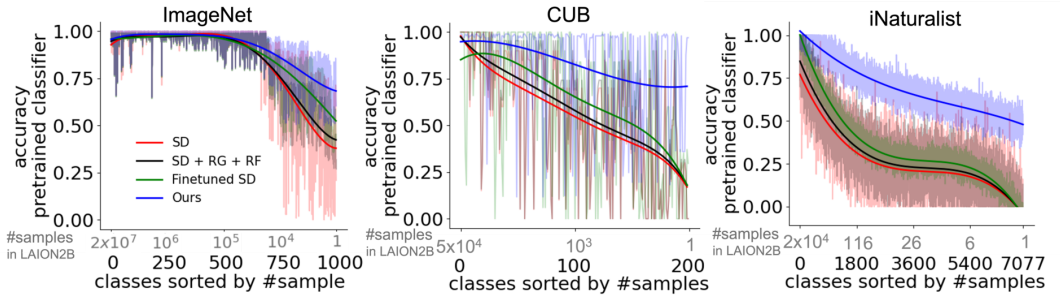

Figure 4: Per-class accuracy of pre-trained object recognition given images generated using various approaches. Classes are ordered by their number of occurrences in LAION2B. SeedSelect achieves the highest accuracy for all classes across all benchmarks, outperforming previous methods. Corresponding tables can be found in Supp. Solid lines: Polynomial fits.

METHOD	FID ↓
SD (ROMBACH ET AL. 2022)	**6.4**
SD+RG+RF (HE ET AL. 2023)	6.9
FINETUNED SD (AZIZI ET AL. 2023)	10.2
SEEDSELECT (OURS)	**6.5**

Table 1: Quality of images generated using various approaches, measured using FID. Compared with SD, other methods for rare-concept generation hurt image quality, but SeedSelect maintains the same image quality as SD.

Human Eval	**ImageNet** Many #>1M	Med 1M>#>10K	Few 10K>#
Finetuned SD	48.01±1.01	41.55±1.55	15.84±2.29
SeedSelect	**50.12±1.00**	**55.33±1.42**	**69.08±2.46**
Neither	1.87±1.12	3.12±1.48	15.08±2.22

Human Eval	**CUB** All	**iNaturalist** All
Finetuned SD	20.18±2.31	14.45±2.77
SeedSelect	**68.98±2.71**	**72.44±2.13**
Neither	10.84±3.11	13.11±2.79

Table 2: Human evaluation for rare-concept generation. Values are the percentage of raters that selected each option.

6.1 Rare Concept Generation

We evaluate the quality of images from rare classes generated by our approach.

Datasets. We evaluated SeedSelect on three common benchmarks: **(1) ImageNet (Deng et al. 2009):**, the cannonical dataset with 1000 classes. As shown in Figure 2, about 25% of ImageNet classes are in the tail of Laion. **(2) CUB (Wah et al. 2011):** A *fine-grained* dataset with a total of 200 bird categories. Most of the classes are in the tail of the Laion distribution. **(3) iNaturalist (Van Horn et al. 2018)**: A large-scale, *fine-grained* dataset for species classification. Its entire set of classes is in the tail of the Laion distribution.

We use CUB and iNaturalist since most of their classes are rare; i.e. have been represented by fewer than 10k samples in the training set of the diffusion model (more details in supp).

Evaluation protocol. We ranked classes for each dataset according to their occurrence frequency in the LAION2B dataset. For each class, the set of reference images for all methods was taken from the trainset. Specifically, we sampled a maximum of 50 random images, $k = \max(|class|, 50)$. Subsequently, we used different methods to generate images based on the real reference samples.

Pretrained classifiers. To measure the correctness of generated images, we use SoTA pre-trained classifiers for each benchmark, sourced from open repositories available online. Specifically, for ImageNet, we used (Tu et al. 2022), achieving 88.2% accuracy on the corresponding test set. For CUB, we used (Chou, Kao, and Lin 2023), which attains 93.1% test accuracy. For iNaturalist, we used (Ryali et al. 2023), which has 83.8% test accuracy.

Compared Methods. We compared SeedSelect with the following methods. **SD (Rombach et al. 2022):** Vanilla Stable Diffusion v2.1; **SD+RG+RF (He et al. 2023):** Stable Diffusion v2.1 with Real Guidance (RG) and Real Filtering (RF).; and **Finetuned SD (Azizi et al. 2023):** Finetuning SD on real training samples for each class. To ensure a fair comparison, we replicated the methods mentioned above with StableDiffusion v2.1 using the code published by the respective authors. This step was essential because the previous approaches relied on different versions of pre-trained text-to-image models, making it crucial to establish a consistent framework for evaluation.

Generation Protocol: See supplementary material.

Evaluate faithfulness using pre-trained classifiers. Figure 4 shows per-class accuracy of different generation approaches on different benchmarks, as evaluated by pre-trained classifiers. Classes are ranked by their prevalence in the LAION2B dataset. Notably, it shows that while existing methods falter in generating less common semantic concepts, our approach consistently achieves higher accuracy across all classes within all benchmarks.

Evaluating realism and visual appeal. Table 1 further presents SeedSelect image quality in terms of realism and visual appeal compared to current generation methods. The

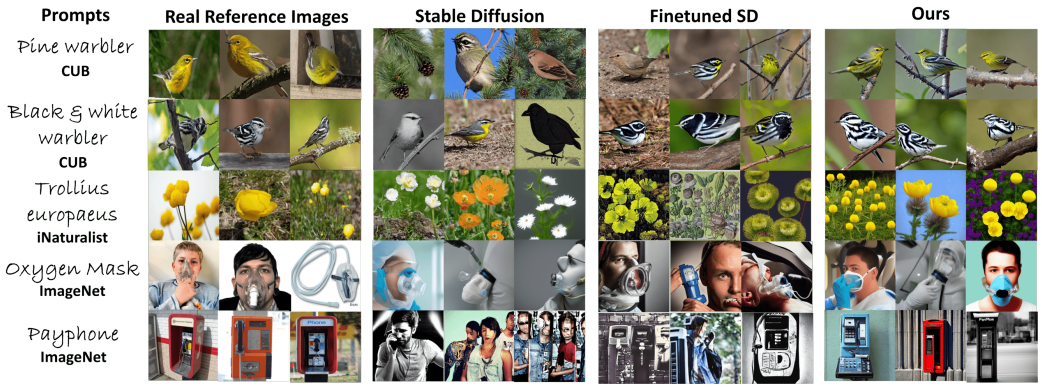

Figure 5: Qualitative comparison. Images generated by various methods for 5 rare classes from 3 datasets. Images generated by the competing techniques may exhibit high quality but frequently contain inaccuracies and fail to align with the real concept.

measurement of this quality is done using the FID, which was calculated between 50K generated and 50K real ImageNet test images. The results demonstrate that SeedSelect's capability of generating rare concepts is not traded with image naturalism. While other methods are negatively affected by adaptation to rare-concept generation, SeedSelect can generate images with the same quality as vanilla SD. This is attributed to the fact that these methods fine-tune the diffusion model or modify its denoising process, leading to a decline in image realism. In contrast, our method uses a pre-trained diffusion model with fixed parameters, only optimizing its seed during the generation process.

Qualitative analysis. Figure 5 compares images generated by Stable Diffusion and Finetuned SD (Azizi et al. 2023) with our approach on rare concepts from CUB, ImageNet, and iNaturalist. See Suppl. for additional examples. The results show that although the compared methods generate realistic images of high quality they often fail to generate the correct concept.

Evaluation with human raters. We further performed a user study to analyze the correctness of the generated images. We randomly selected 30 classes from CUB, and iNaturalist, and 90 classes from ImageNet (30 for head, 30 for med, and 30 for tail). For each class we generated 10 images with SeedSelect and Finetuned SD (Azizi et al. 2023), the best baseline found in the previous analysis. Respondents were given the class name, three real samples as a reference, and the two generated images. They were asked to select which generated image better fits the class name and is semantically similar to the reference images. The final score for each approach is calculated as the number of times respondents selected the approach, averaged across all the classes in the set. The study results are shown in Table 2. SeedSelect received the highest percentage of votes across all benchmarks: it is $\times 3.4$ better on CUB and $\times 5$ on the iNaturalist. Moreover, it excels across all splits of ImageNet with the most notable advancement observed within the tail, achieving a $\times 4.3$ increase in accuracy. These results are correlated with the classifier results in Figure 4.

	SD	SD+RG+RF	FINTUNED SD	OURS
NDB ↓	2.48	2.6	2.9	2.52
PRECISION ↑	0.79	0.70	0.61	0.77
RECALL ↑	0.2	0.15	0.11	0.18
FIDELITY ↑	0.85	0.79	0.71	0.83
DIVERSITY ↑	0.37	0.28	0.20	0.36

Table 3: Diversity analysis comparison. SeedSelect generated samples with high diversity as SD, while other approaches hurt diversity.

Diversity Analysis. A potential concern that may arise pertains to the diversity of the images generated by our approach. We analyze the diversity of images generated by SeedSelect compared to current methods using two measures of diversity. First, using NDB which finds diversity through mode collapse analysis, (Richardson and Weiss 2018). Second, diversity as measured by (Naeem et al. 2020), which directly assesses the coverage of generated samples compared to real samples. We also report Precision, Recall, and Fidelity. These metrics were computed using 50K generated vs. 50K real ImageNet test images. We determined hyperparameters such as the number of clusters or neighbors, using the Elbow method (Thorndike 1953).

The results of the diversity analysis are presented in Table 3, indicating that SeedSelect maintains similar diversity to SD, whereas competing methods show lower diversity. We attribute this result to the fact that SeedSelect is initialized with a random seed and then optimized, leading to distinct images for each seed.

Generation time. We compare our approach with personalized generation methods. Both Textual-Inversion (Gal et al. 2023a) and Dreembooth (Ruiz et al. 2023) require 30-60 minutes to learn a new single concept on a single NVIDIA A100 GPU. In contrast, SeedSelect with bootstrapping (See Section 5.1) takes 1-5 minutes to adapt to the new concept and 1-2 seconds to generate new semantically correct images.

	Rater decisions		
	Stable Diffusion	SeedSelect	Neither
Matches prompt	16.22±2.6	**70.21±2.6**	13.57±4.1
Looks realistic	16.19±5.8	**62.46±6.5**	21.35±7.2

Table 4: Human evaluation of hand-palm generated images. Values are percentage of raters that selected each option.

6.2 Hand Generation

As a first application, we test SeedSelect on hand generation. Generating well-formed hand palms has been infamously hard to achieve with diffusion models (Zhang and Agrawala 2023). We tested how SeedSelect can be used for improving generation of hand palms. Since there are currently no standard benchmark or automated methods to evaluate the quality of hand palm generation, we evaluated the results by asking human raters. In short, we used a 2-alternative-forced choice design (2AFC) asking raters to select if they prefer an image generated by SeedSelect or by SD. The detailed procedure of the experiment is in the supplemental material.

Table 4 shows the results of the user study. SeedSelect is $\sim \times 4.5$ better in matching the prompt and $\sim \times 4$ in generating realistic hands. More in supplementary. Figure 6 compares Stable Diffusion with our approach on 5 hand prompts.

6.3 Synthetic Data for Few-Shot Recognition

We further examine the advantages of using SeedSelect for few-shot classification through semantic data augmentation.

In the context of few-shot image recognition, we are provided with a limited number of real training samples per class, along with their corresponding class names, and the goal is to fine-tune CLIP.

Experimental Setup. For a fair comparison we follow the same experimental protocol of (Zhou et al. 2021; Zhang et al. 2022) and generation protocol of (He et al. 2023). Specifically, given a limited number of real training samples per class we generated 800 samples for each class using SeedSelect. We then fine-tuned a pre-trained CLIP-RN50 (ResNet-50). Fine-tuning is done using both real and generated images known as mix-training. More details on the setup/protocol can be found in (He et al. 2023) and in Supp.

Compared Methods. We compare our approach with the following baselines: **Zeor-shot CLIP:** Applying the pre-trained CLIP classifier without fine-tuning; **CooP (Zhou et al. 2021):** Fine-tuning a pre-trained CLIP via learnable continuous tokens while keeping all model parameters fixed; **Tip Adapter (Zhang et al. 2022):** Fine-tuning a lightweight residual feature adapter; **CT & SD:** Classifier tuning with images generated with SD. **Textual Inversion (Gal et al. 2023a):** Classifier tuning with images generated using personalized concepts (See Supp for implementation detail). Results for **CT & SD** were reproduced by us on SDv2.1 using the code published by the respective authors.

Results Figure 7 shows results for the few-shot image recognition task. It demonstrates the effectiveness of SeedSelect in generating high-quality augmentations. When using SeedSelect augmentations to fine-tune a CLIP classifier, we achieve state-of-the-art performance across all shots. Remarkably, even with just a single image for training, SeedSelect still manages to produce valuable, diverse, and better augmentations compared to previous baselines. We provide results for iNaturalist in the Supp material.

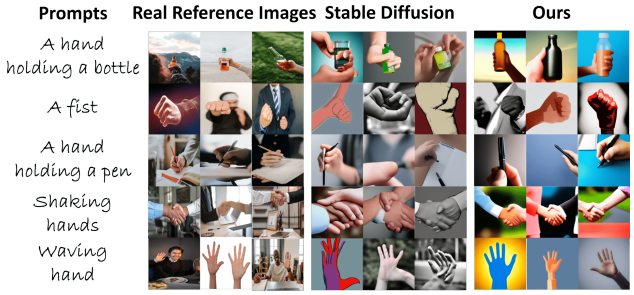

Figure 6: Qualitative comparison. Images generated by Stable Diffusion and SeedSelect for several hand generation prompts. While generated hands from SD are often corrupted SeedSelect can fix this shortcoming given a few reference examples.

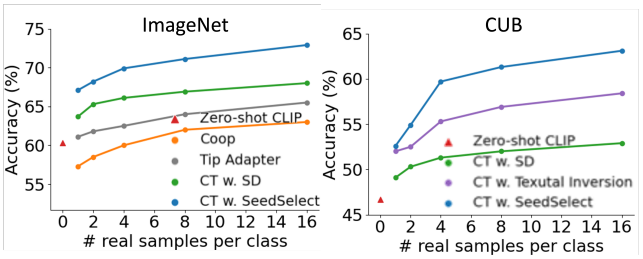

Figure 7: Results for few-shot image recognition, comparing SeedSelect to previous approaches.Fine-tuning a CLIP classifier on SeedSelect generated images consistently achieves SOTA results across all shot levels, with SeedSelect performing well even when given just a single image.

7 Discussion and Limitations

Although very powerful, modern text-to-image generation models still suffer from several shortcomings. They often generate incorrect images when prompted for rare concepts especially when a closely related concept appears frequently in the train set of the diffusion model. We propose to remedy these issues by providing a handful of reference images of the concept to the diffusion model. Essentially, it selects a generation seed that drives the diffusion model to generate the correct concept, semantically and visually. While SeedSelect is simple there are several limitations to consider. First, we find that it struggles with imitating the style of the reference images (e.g. when guided by sketch images of dogs, SeedSelect often generates natural images of dogs rather than sketches). Second, the optimized z_T is prompt-specific, and doesn't generalize directly to other prompts. Finally, for extremely rare concepts that have only few examples in LAION2B, the quality of generated images is poor.

References

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. SpaText: Spatio-Textual Representation for Controllable Image Generation. *CVPR*.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic Data from Diffusion Models Improves ImageNet Classification. *ArXiv*.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *SIGGRAPH*.

Chou, P.-Y.; Kao, Y.-Y.; and Lin, C.-H. 2023. Fine-grained Visual Classification with High-temperature Refinement and Background Suppression. *ArXiv*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai Li; and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *cvpr*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. *ICLR*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023b. Designing an Encoder for Fast Personalization of Text-to-Image Models. *arXiv preprint arXiv:2302.12228*.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P. H. S.; Bai, S.; and Qi, X. 2023. Is synthetic data from generative models ready for image recognition? *ICLR*.

Ho, J.; and Salimans, T. 2021. Classifier-free diffusion guidance. *NeurIPS workshop on Deep Generative Models and Downstream Applications*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *NeurIPS*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*.

Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655*.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *ECCV*.

Liu, V.; and Chilton, L. B. 2022. Design guidelines for prompt engineering text-to-image generative models. In *ACM SIGCHI*.

Marcus, G.; Davis, E.; and Aaronson, S. 2022. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*.

Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. Proceedings of Machine Learning Research. PMLR.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *NeurIPS*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rassin, R.; Hirsch, E.; Glickman, D.; Ravfogel, S.; Goldberg, Y.; and Chechik, G. 2023. Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. *ArXiv*.

Richardson, E.; and Weiss, Y. 2018. On gans and gmms. *NeurIPS*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*.

Ryali, C. K.; Hu, Y.-T.; Bolya, D.; Wei, C.; Fan, H.; Huang, P.-Y. B.; Aggarwal, V.; Chowdhury, A.; Poursaeed, O.; Hoffman, J.; Malik, J.; Li, Y.; and Feichtenhofer, C. 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. *ICML*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *SIGGRAPH*.

Thorndike, R. L. 1953. Who belongs in the family? *Psychometrika*.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MaxViT: Multi-Axis Vision Transformer. *ECCV*.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018.

The inaturalist species classification and detection dataset. In *CVPR*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. *arXiv preprint arXiv:2210.14896*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y. J.; and Li, H. 2022. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *ECCV*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2021. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2021. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*.