Diffusion Curriculum: Synthetic-to-Real Data Curriculum via Image-Guided Diffusion

Supplementary Material

A. Appendix

A.1. Motivation for DisCL's Data Selection

When curating data for a training curriculum, real data often aligns with the test distribution better but suffers from deficiency, noise, low quality, or imbalance; Synthetic data can potentially fix these problems but suffers from a large distribution gap to the test. Our synthetic-to-real curriculum is designed to **combine the complementary strengths of both data types and overcome their weaknesses**. Unlike previous methods using synthetic data with no real-image guidance or a fixed guidance level, DisCL dynamically adjusts the real-image guidance level per training stage to generate a spectrum of synthetic-to-real samples that accelerate learning progress and meanwhile progressively bridging the distribution gap. Unlike pre-defined easy-to-hard curricula on real data, DisCL's data selection is adaptive to the training dynamics, considers diversity and distribution gap, and is optimized for achieving the greatest progress per stage.

A.2. Synthetic Data Generation with Image Guidance

In this section, we visualize more generated images in (Phase 1) of our method with various levels of image guidance, for two different classification tasks.

A.2.1. Generation Settings and Statistics

We provide the statistics for the synthetic data generation within our paradigm on ImageNet-LT, CIFAR100-LT, iNaturalist2018, and iWildCam, as shown in Table 5.

A.2.2. ImageNet-LT Synthetic Generation

Selection of Text prompts To improve model performance on the minority classes, high-quality and diverse synthetic samples are required. To achieve so, we follow the approach in Fu et al. [10], and utilize publicly available GPT-3.5-turbo to generate diverse prompts for these 1000 IN-LT classes. We use the following prompt to query GPT-3.5-turbo for generating descriptions for class X:

"*Please provide 10 language descriptions for random scenes that contain only the class X from the ImageNet-LT dataset. Each description should be different and contain a minimum of 15 words. These descriptions will serve as a guide for Stable Diffusion in generating images.*"

The sample-prompts generated by GPT-3.5-turbo are listed in Table 6.

Selection of Images Guidance Levels We first analyze the cosine similarity between synthetic images and real images, as well as between synthetic images and text prompts. The similarity score between synthetic images and real images can be used to quantify the diversity introduced in the synthetic images. As depicted in Fig. 6a, the similarity between synthetic images and real images decrease as the guidance level reduces, demonstrating the trend of increased diversity in the data spectrum. However, the changes in the scores are relatively small across varying guidance levels. Combined with the visual cases for this dataset (examples shown in Fig. 8), we observe that for images generated with high guidance levels ($\lambda \geq 0.7$), only minor details are modified by the diffusion model, resulting in high similarity scores above 0.85. However, we aim to provide more diverse synthetic data to increase the model's generalization on the class-balanced test set. Including these highly similar images may hinder the diversity and cause the model to overfit to specific visual features, thereby negatively impacting its generalization ability. Therefore, we select $\{0.0, 0.1, 0.3, 0.5\}$ as the interval of image guidance levels used in the training process for this dataset.

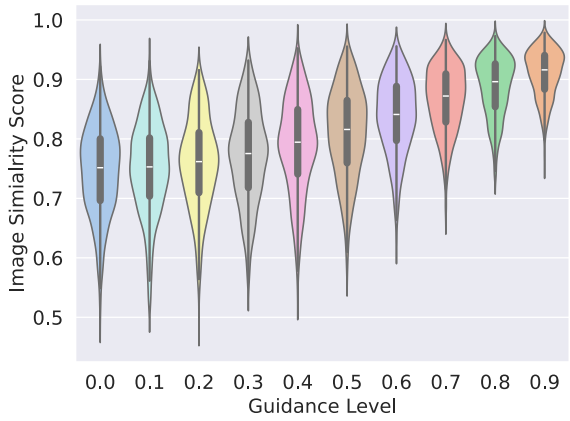
Selection of CLIPScore Threshold We leverage the widely used CLIPScore [18] to filter out poor-quality images in the synthetic data. In this method, the CLIP cosine similarity between synthetic images' embeddings and text embeddings is computed to measure the alignment between images and the corresponding classes provided in text prompts. For the synthetic data generation for ImageNet-LT, we use a unified template that emphasizes the class information in text prompts. Following Trabucco et al. [38], we use "*a photo of <class name>*" to prompt the CLIP model and compute the cosine similarity. We also consider the value of the filtering threshold for synthetic data. Following previous work [33], we set the threshold to 0.3 based on the distribution of similarity scores and a review of generation quality, as shown in Fig. 6b. We observe that a threshold of 0.3 effectively filters out synthetic images with poor quality or mismatched classes.

A.2.3. iWildCam Synthetic Generation

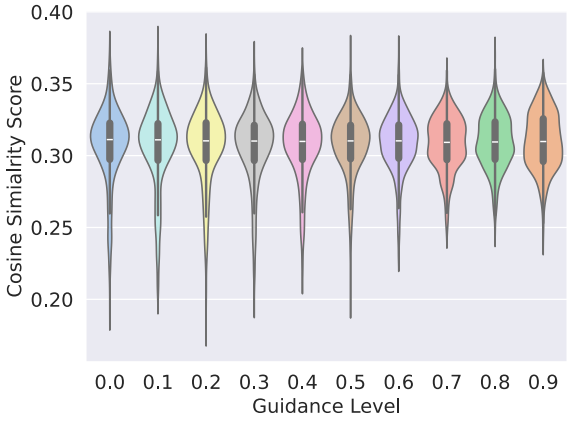
Selection of Text prompts Following previous work [8, 38], we first define prompts for each class using the template "*a photo of <class>*". However, the classnames in iWildCam comprises of scientific names, which are usually unseen/unknown concepts to the diffusion text encoder. For example, "canis lupus" is the class name for "wolf" animal. To address this, we replace the scientific names with their common names and add a postfix "*in the wild*" in the prompt to drive the generation of wild images. The final text prompt we use is "*a photo of <common name of class> in the wild*".

Images' Details	ImageNet-LT	CIFAR100-LT		iNaturalist2018	iWildCam
		Irb=100	Irb=50		
No. of Hard Samples	1643	324	268	44956	8260
Number of Image Guidance Scales λ	4	4	4	4	3
Number of Random Seed Per Image	8	8	8	4	8
Number of Generated Images	51917	2592	2144	179824	197756
Number of Generated Images After Filtering	24141	809	668	75234	90093
Acceptance Rate	46.50%	31.21%	31.16%	41.84%	45.56%

Table 5. Statistics about Generated Synthetic Data. Irb refers to the imbalance ratio used to sample CIFAR100-LT dataset.



(a) Similarity b/w synthetic images & its original real image.



(b) Similarity b/w synthetic images & defined text prompt.

Figure 6. CLIP Cosine similarity score for ImageNet-LT Synthesis.

Selection of Images Guidance Levels Based on the generated data with multiple image guidance scales, we search for effective image guidance scales for this task using CLIP cosine similarity scores between synthetic image embeddings and real image embeddings. As shown in Fig. 7a, as the difference between real images and synthetic images increases, the cosine similarity between image embeddings decreases from $\lambda = 1$ to $\lambda = 0.3$. However, when the image guidance continues to decrease to $\lambda = 0$, the cosine similarity score increases slightly. With low image guidance scales, the diffusion model tends to generate images that heavily rely on text information, maintaining only global information (such as the color of the image background) in the synthetic data for some images. This creates a distribution gap between these synthetic data and real data that is too large for the model to accurately compare the differences between the two images using embedding representation. Additionally, based on the analysis of the quality of synthetic images and to leverage the difficulty of the features and the distribution gap between synthetic and real data, we set the image guidance scales to $\{0.5, 0.7, 0.9\}$ for this task.

Selection of CLIPScore Threshold To filter out low-quality images, we assess the CLIP cosine similarity scores between synthetic image embeddings and corresponding text embeddings for each class. We use the same prompt template as in the generation process ("*a photo of <common name for animal> in the wild*") to compute CLIPScore for synthetic images. The distribution of CLIPScores is shown in Fig. 7b, which reveals a distinct gap around 0.25. Combined with a review of the quality of synthetic data, we set the threshold to 0.25. Synthetic data with a CLIPScore lower than 0.25 are considered poor-quality samples.

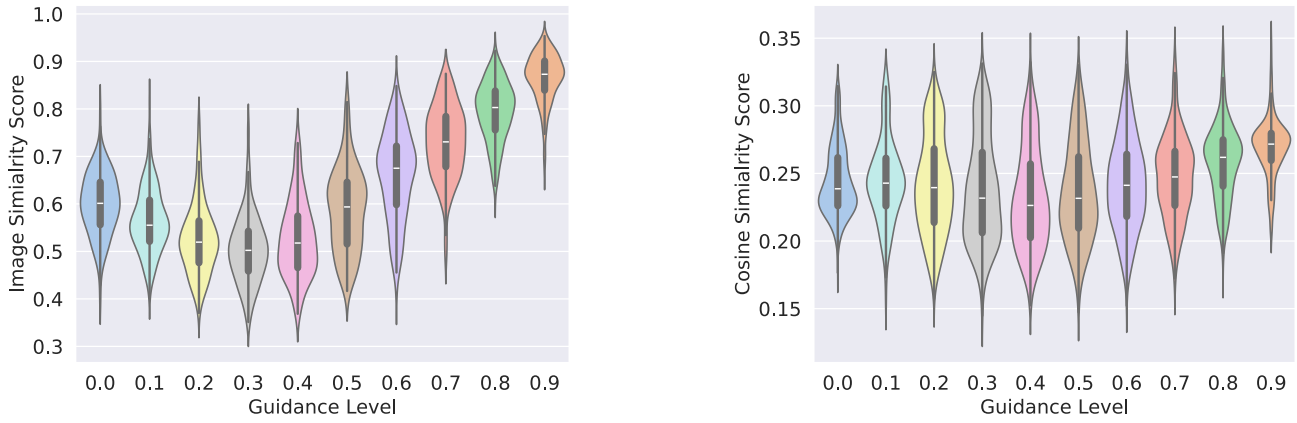
A.2.4. Visualization

Visual Cases We provide additional visual examples of synthetic data generated with multiple guidance levels and text prompts for the ImageNet-LT and iWildCam datasets. The results are visualized in Fig. 8 and Fig. 9. These examples demonstrate that the model can generate synthetic data with various postures, backgrounds, and actions as the image guidance level decreases. Particularly for ImageNet-LT generation results, diverse prompts introduce more varied features into low-guidance data. These diverse features enable the model to achieve better generalization on the target distribution.

Failure Cases During generation, despite designing text prompts and applying CLIPScore to filter to remove low-quality data, some failure cases still occur in the synthetic dataset. In this section, we discuss these failure cases encountered during the generation process. As shown in Fig. 10 and Fig. 11, the first failure case is caused due to the in-

Class Name	Prompts
Grand Piano	A grand piano sits elegantly in a sunlit room, its glossy finish reflecting the warm glow. In a cozy living room, the grand piano adds a touch of luxury and sophistication to the space. The grand piano sits silently in a dimly lit room, waiting patiently for a skillful pianist to bring it to life. In a grand ballroom, the grand piano provides a majestic backdrop for a glamorous event. A vintage grand piano exudes timeless elegance in a quaint parlor, filled with antique charm.
Pufferfish	A colorful pufferfish swimming gracefully in a crystal-clear ocean, surrounded by vibrant coral reefs. A group of playful pufferfish blowing bubbles and chasing each other in a sunlit underwater cave. A shoal of pufferfish moving in unison, creating a mesmerizing dance of synchronized swimming in the deep sea. A fierce pufferfish defending its territory from intruders, puffing up its body and displaying its sharp spikes as a warning. A baby pufferfish following its larger parent closely, learning the ropes of survival in the vast ocean ecosystem.

Table 6. Generated text prompts for ImageNet-LT classes



(a) Synthetic image & original real images.

(b) Synthetic image & defined text prompt.

Figure 7. CLIP Cosine similarity score for iWildCam Synthesis.

ability to recognize objects in the original images. If these objects are clearly obscured or hard-to-identify (e.g. second case in Fig. 11 and first case in Fig. 10), diffusion models cannot accurately identify the object or modify details for generating diverse and useful data. For these seed images, only synthetic data generated with a low-guidance scale can achieve a CLIPScore higher than the threshold. However, this approach compromises the smooth transition of data from synthetic to real distribution. Even though the diffusion model can generate images with a smooth transition for most-of-the-cases, our quality-check on synthetic data can constrain the feature extraction and alignment ability of the CLIP model. For example, in second case of Fig. 10, CLIPScore filters out the slightly modified but perceptually useful images, containing prototypical class features.

A.3. Application of DisCL to Other Datasets and Model Scale

To further assess the robustness of DisCL, we extend our experiments to two additional widely used imbalanced datasets: CIFAR-100-LT [6] and iNaturalist2018 [40]. For iNaturalist2018, we generate synthetic data following the same approach and settings used for the long-tail classification task on ImageNet-LT. In the case of CIFAR-100-LT dataset, due to the lower image resolution, we adjust the image guidance scale to $\{0.5, 0.7, 0.9\}$ so as to ensure high-quality synthetic data generation. Visual examples of the generated data are shown in Fig. 12 and 13. For CIFAR-100-LT, we evaluate

Figure 8. Synthetic generation with various image guidance and random seeds based on ImageNet-LT.

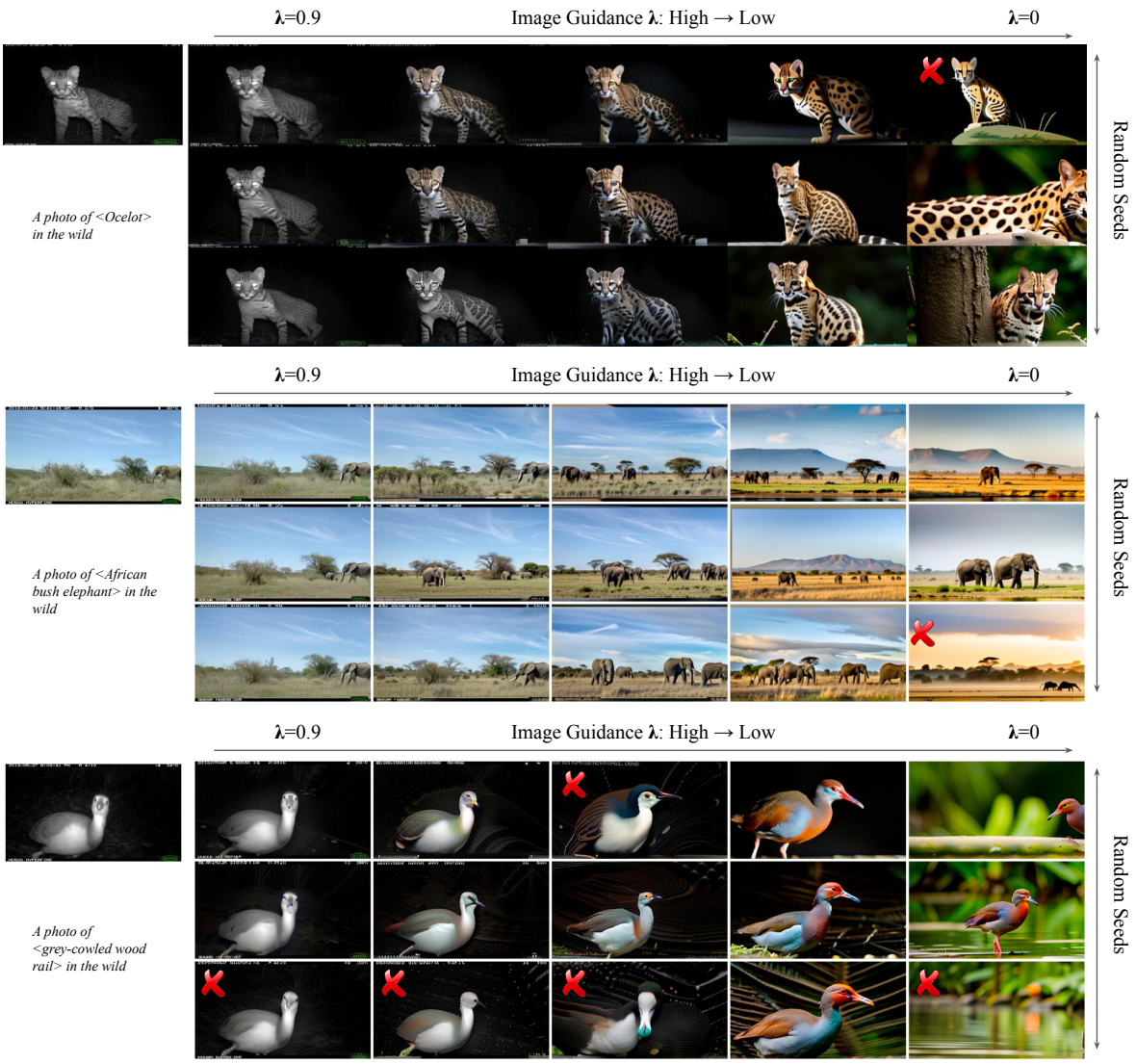


Figure 9. Synthetic generation with various image guidance and random seeds based on iWildCam.

the performance of DisCL under different imbalance ratios (50 and 100). Additionally, we expand our model evaluation to a larger scale, ResNet-34 (widely adopted for ImageNet) with the same experimental settings of DisCL as before. As evident from Table 2 and Table 3, our results demonstrate that DisCL achieves a notable improvements in overall top-1 accuracy (e.g., +1–3.3% over baselines) and few class performance (e.g., +3-8% for tail classes) across both datasets. We also notice that combining a class-reweighting loss (BS) with DisCL causes an oversaturation in tail-class signals, causing the model to neglect *many* classes during the training. This suggests that reweighting and mixing synthetic with real data address different aspects of class imbalance; aligning with prior works, [1] and [35]. *Notably*, the top-1 accuracy gains persist when scaling the model to ResNet-34, as demonstrated for CIFAR-100-LT in Table 7 and ImageNet-LT in Table 8. This underscores the flexibility of our proposed DisCL method across different datasets and model scales.

A.4. Training with Curriculum Learning

A.4.1. Long-Tail Learning with Non-Adaptive Strategy

For long-tail classification, we propose a non-adaptive curriculum learning strategy that starts with the lowest guidance and progressively increases to the highest guidance within the defined interval Λ. We employ a linear scheduler to adjust the guidance levels during training, allowing the model to train with data from various guidance levels for *equal durations*. Furthermore, the test set of ImageNet-LT is in-distribution to its training data; unlike the training data, it is a class-balanced set. To mitigate the potential negative effects of the distribution gap between synthetic and real data, all

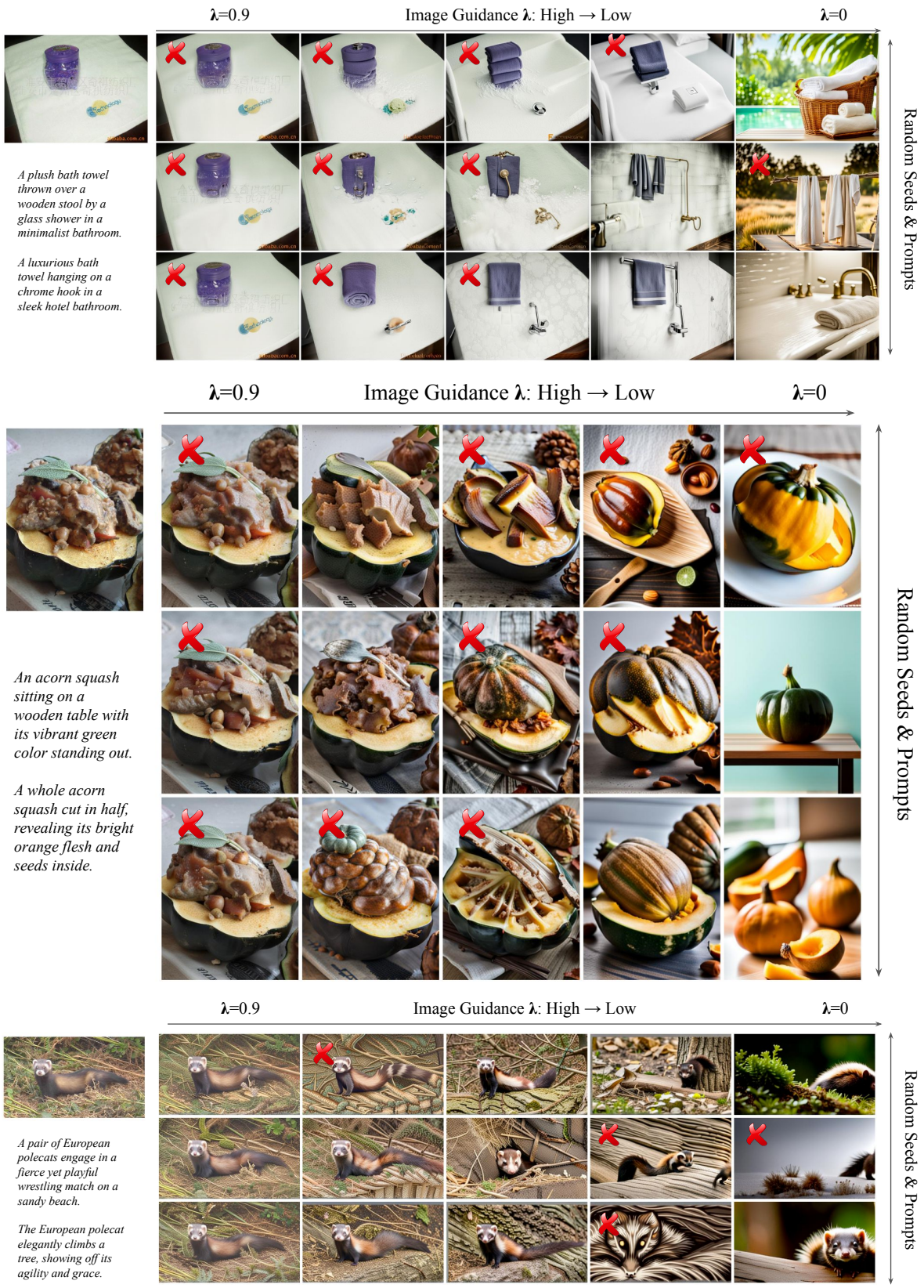


Figure 10. Failure cases for ImageNet-LT synthetic generation

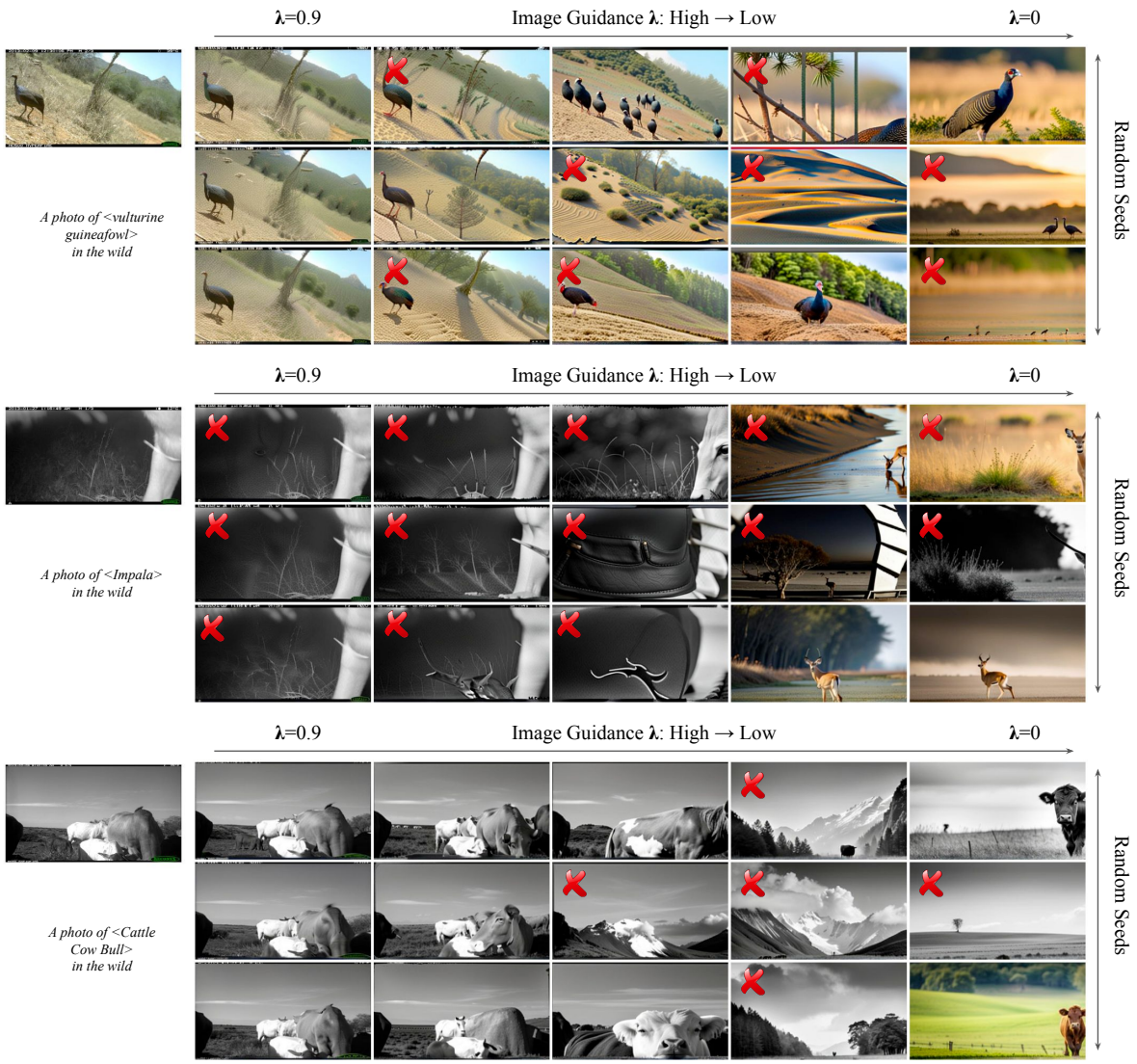


Figure 11. Failure cases for iWildCam synthetic generation

Method	Curriculum	CIFAR-100-LT (Imbalance Ratio=100) Many	Medium	Few	Overall	CIFAR-100-LT (Imbalance Ratio=50) Many	Medium	Few	Overall
CE	N/A	**51.71**	23.51	5.05	27.7	**52.14**	**29.97**	10.7	32.04
CE + DisCL	Diverse to Specific	49.83	23.26	**7.9**	**28.4**	51.83	29.12	**12.64**	**32.18**
BS	N/A	**46.23**	**28.0**	13.13	29.79	**46.48**	**33.48**	22.1	34.6
BS + DisCL	Diverse to Specific	44.9	27.4	**16.8**	**30.3**	45.51	32.08	**23.99**	34.5

Table 7. Accuracy (%) of ResNet-34 on CIFAR-100-LT classification task with imbalance ratios of 100 and 50, highlighting the best accuracy in bold for overall and class categories (*many*, *medium*, and *few*).

the hard tail samples from original data are involved into training at all times. Furthermore, with DisCL, number of samples for tail classes increases along with the introduction of synthetic data at each stage, however the ratio of tail-to-nontail samples is still very skewed. To preserve a constant imbalance-ratio throughout all training stages and experiments, we undersample the non-tail samples at "each stage" so that ratio of tail-samples to non-tail samples matches the proportion of tail classes to non-tail classes present in the original data (13.6%).

All experiments are conducted based on this proportion setting. Complete strategy details are covered in Algorithm 1.

Method	Curriculum	ImageNet-LT Many	Medium	Few	Overall
CE	N/A	63.01	35.90	10.10	42.98
CE + CUDA	N/A	62.78	36.91	11.92	43.34
CE + DisCL	Diverse to Specific	**63.54**	**36.93**	**13.64**	**44.26**
BS	N/A	**62.78**	36.91	11.92	43.34
BS + CUDA	N/A	57.16	44.5	30.49	47.33
BS + DisCL	Diverse to Specific	58.82	**45.21**	**32.53**	**48.42**

Table 8. Accuracy (%) of ResNet-34 on ImageNet-LT classification task, highlighting the best accuracy in bold for overall and class categories (*many*, *medium*, and *few*).

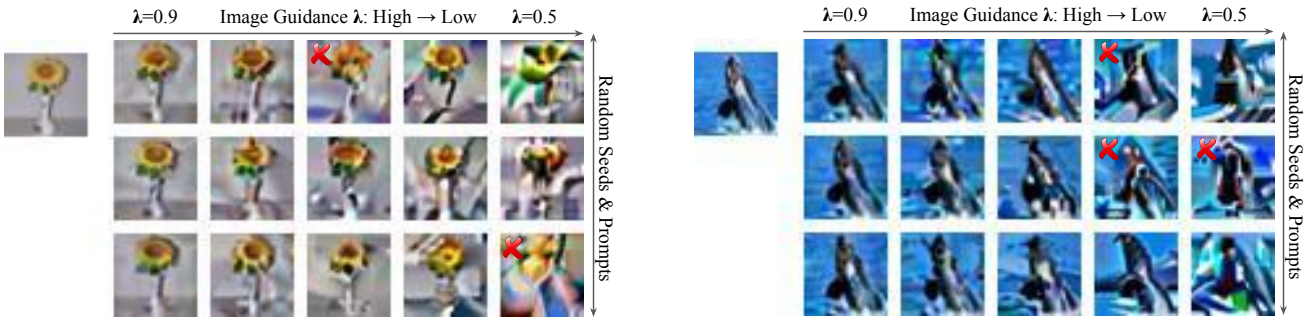


Figure 12. Synthetic generation with various image guidance and random seeds based on CIFAR100. Sample Prompt: *(1) A bright sunflower standing tall in a field, basking in the warm sunlight of a summer day. (2) A majestic whale breaches the surface of the deep blue ocean, sending a spray of water into the air.*

A.4.2. Learning from Low-Quality Data with "Adaptive Curriculum" Strategy

An approximation method to assess the effectiveness of samples in helping model achieve greatest progress on and fastest learning face is introduced by DoCL [46] as shown in Eq 4.

$$\begin{aligned} &\mathbb{E}_{x\in D, x\sim\mathcal{D}}\langle y - f(x), \frac{\partial f(x)}{\partial t}|_S\rangle \\ &\approx \frac{1}{|D|}\sum_{j\in\mathcal{V}}\langle y^{(j)} - f(x^{(j)}), \frac{\partial f(x^{(j)})}{\partial t}|_D\rangle \end{aligned} \tag{4}$$

where $\mathcal{D}$ is the training distribution and $x \in D$ is a set of finite samples randomly sampled from the original distribution $\mathcal{D}$. $\mathcal{V}$ denotes the subset of samples from S. Here, y and $f(x)$ denotes the target-class and sample prediction. $\langle y - f(x), \frac{\partial f(x)}{\partial t}|_{\mathcal{V}}\rangle$ represents the project of residual $y - f(x)$ on the model dynamics $\frac{\partial f(x)}{\partial t}|_{\mathcal{V}}$. This equation indicates that when trained with subset $\mathcal{V}$, the expected progress $\mathbb{E}$ of samples in the original training dataset can be approximated by the progress of samples on subset $\mathcal{V}$ achieved via training on the set D.

For learning from low-quality data, we adopt DoCL and implement an adaptive curriculum strategy to select the synthetic data with best guidance level for each training stage. We showcase the implementation in Algorithm 2, wherein we preserve i for indexing the guidance level in Λ and j for indexing the sample in a given dataset. Before the training process, we randomly select samples from the spectrum for each guidance level in Λ and mark it as guidance validation set $\mathcal{V}$ for progress evaluation. This set has zero overlap with the training data $\mathcal{D}_{\text{all}}$. At each training stage, we randomly sample a set D (termed as random-real set) from the training

Algorithm 1: Training with **non-adaptive** curriculum strategy

Input: Image guidance levels $\Lambda = \{\lambda_i \mid \lambda_i \in [0, 1]\}$,
Non-hard samples $\mathcal{D}_{\text{nh}} = \{(x^{(j)}, y^{(j)}, \lambda^{(j)} = 1)\}_{j=1}^{N}$,
Spectrum of syn-to-real data
$\mathcal{S} = \{(x'^{(j)}, y^{(j)}, \lambda^{(j)}) \mid \lambda^{(j)} \in \Lambda\}_{j=1}^{M}$,
Original hard samples
$\mathcal{D}_{\text{h}} = \{(x^{(j)}, y^{(j)}, \lambda^{(j)} = 1) \mid (x^{(j)}, y^{(j)}, \lambda^{(j)}) \in \mathcal{S}\}$,
Total training epochs E, curriculum cutoff E_{CL},
Predefined linear guidance schedule
$\mathcal{G} = \{\lambda_1, \lambda_2, \ldots, \lambda_e, \ldots, \lambda_{E_{CL}}\}$
Output: Trained model f_θ
Initialize: Pretrained model f_θ

for $e \leq E_{CL}$ **do**
  $\lambda_e = \mathcal{G}(e)$
  Extract $\mathcal{S}_{\lambda_e} = \{(x'^{(j)}, y^{(j)}, \lambda^{(j)}) \mid \lambda^{(j)} = \lambda_e\}$
  Gather new training set $\mathcal{D}_e = \mathcal{S}_{\lambda_e} \cup \mathcal{D}_{\text{nh}} \cup \mathcal{D}_{\text{h}}$
  Finetune model f_θ with $\mathcal{D}_e$
end
for $E_{CL} < e \leq E$ **do**
  Gather new training set $\mathcal{D}_e = \mathcal{D}_{\text{nh}} \cup \mathcal{D}_{\text{h}}$
  Finetune model f_θ with $\mathcal{D}_e$
end

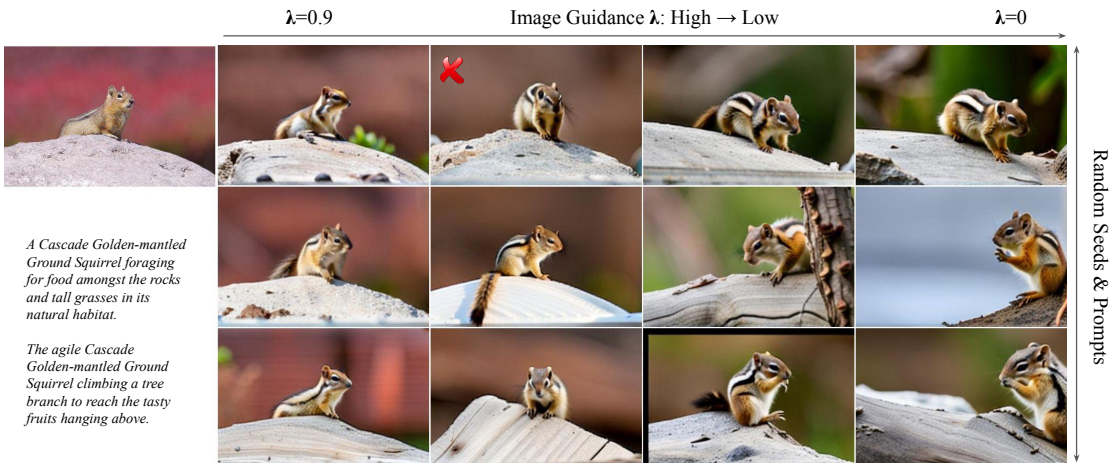


Figure 13. Synthetic generation with various image guidance and random seeds based on iNaturalist 2018.

dataset $\mathcal{D}_{\text{all}}$. Before selecting the guidance level, we train the model on dataset D and evaluate the progress (in terms of classifier's prediction score) achieved on samples of each subset $\mathcal{V}_i$ corresponing to a given guidance λ_i. We then select the λ_i with the highest progress to gather synthetic data and combine it with other non-hard samples from the original training data for the current training stage. This technique encourages the model to adaptively select the most informative guidance for the current training stage. At the end of the curriculum-training, to alleviate the negative effect of the distribution gap between synthetic data and real data for this task, we keep finetuning the model with real data for a short period. The steps of algorithm are detailed in Algorithm 2.

Algorithm 2: Training with **adaptive** curriculum strategy

Input: Image guidance levels $\Lambda = \{\lambda_i \mid \lambda_i \in [0, 1]\}$,
Non-hard samples $\mathcal{D}_{\text{nh}} = \{(x^{(j)}, y^{(j)}, \lambda^{(j)} = 1)\}_{j=1}^{N}$,
Syn-to-real spectrum data
$\mathcal{S} = \{(x'^{(j)}, y^{(j)}, \lambda^{(j)}) \mid \lambda^{(j)} \in \Lambda\}_{j=1}^{M}$,
Combined training data
$\mathcal{D}_{\text{all}} = \mathcal{D}_{\text{nh}} \cup \{(x'^{(j)}, y^{(j)}, \lambda^{(j)}) \mid \lambda^{(j)} = 1\}$,
Guidance validation set
$\mathcal{V} = \{(x'^{(j)}, y^{(j)}, \lambda^{(j)}) \mid \lambda^{(j)} \in \Lambda\}_{j=1}^{m}$,
Total training epochs E, curriculum cutoff epoch E_{CL},
size of real-random set $|D|$
Output: Trained model f_θ
Initialize: Pretrained model f_θ
`/* Note: `$\mathcal{V}$` has no overlap with `$\mathcal{D}_{\text{all}}$` */`
for $e \leq E_{CL}$ **do**
  Compute true-class probability p_{bef} of model f_θ on $\mathcal{V}$
  Sample a random-real set D from $\mathcal{D}_{\text{all}}$
    `/* contains only real data */`
  Train model f_θ with D
  Compute true-class probability p_{aft} of model f_θ on $\mathcal{V}$
  $\lambda_e \leftarrow \arg\max_{\lambda_i \in \Lambda} (p_{\text{aft}}(\lambda_i) - p_{\text{bef}}(\lambda_i))$
  Extract $\mathcal{S}_{\lambda_e} = \{(x'^{(j)}, y^{(j)}, \lambda^{(j)}) \mid \lambda^{(j)} = \lambda_e\}$
  Form training set $\mathcal{D}_e = \mathcal{S}_{\lambda_e} \cup \mathcal{D}_{\text{nh}}$
  Train model f_θ with $\mathcal{D}_e$
end
for $E_{CL} < e \leq E$ **do**
  Train model f_θ with $\mathcal{D}_{\text{all}}$
end

A.5. Hyperparameters for Synthetic Generation and Model Training

The values of all hyperparameters used for synthetic data generation with diffusion model and curriculum learning strategy are listed in Table 9.

For ImageNet-LT, we implement baselines based on the codebase and the pretrained model from LDMLR. We also re-implement CUDA baseline from this codebase, containing some missing models. We use the same hyper-parameter settings as listed in the CUDA paper. For FLYP, we implement baseline models with FLYP codebase and leverage the available pretrained model from Open CLIP.

A.6. Computational Requirements for Synthetic Generation

For computational requirements of offline generation, 1 RTX A5000 GPU is used to generate synthetic images. For time efficiency, It took 10 seconds to generate a full spectrum (6 image guidance levels) of synthetic images for each real image with resolution=480×270.

A.7. Further Discussion on Experiment Results

In this section, we analyze the results of each guidance level under *Fixed Guidance* experiment to observe the effect of different image guidance levels on the classifier's performance. During the training process, synthetic data generated from only a specific guidance level combined with original real data is presented to the model. The ablation numbers are shown in Fig. 14.

For the iWildCam dataset, data generated with text-only guidance ($\lambda = 0$) has the largest distribution gap between

	Hyperparameter Name	Value
Synthetic Generation	Text Guidance Scale w	10
	Noise Scheduler	DDIM
	Stable Diffusion Denoising Steps	1000
	Stable Diffusion Checkpoint	stabilityai/stable-diffusion-xl-refiner-1.0
	CLIP Filter Model	openai/clip-vit-base-patch32
	Filtering Threshold for iWildCam	0.25
	Filtering Threshold for ImageNet-LT	0.30
	GPU Used	Nvidia rtx5000 with 24GB
ImageNet-LT	Level of Image Guidances λ	$\{0, 0.1, 0.3, 0.5, 1.0\}$
	CLIP Filtering Threshold	0.3
	Batch Size for ResNet-10	128
	Learning Rate	$1e$-3
	Optimizer	Adam
	Scheduler	Cosine
	Training Epoch	65
	Training Epoch for Curriculum Learning	60
	GPU Used	Nvidia rtx5000 with 24GB
iWildCam	Level of Image Guidances λ	$\{0.5, 0.7, 0.9, 1.0\}$
	CLIP Filtering Threshold	0.25
	Size of Dataset D	30000
	Size of Guidance Validate Dataset S	2000
	Batch Size for CLIP ViT-B/16	256
	Batch Size for CLIP ViT-L/16	200
	Learning Rate	$1e$-5
	Optimizer	AdamW
	Scheduler	Cosine with Warmup
	Warmup Step	500
	Training Epoch	20
	Training Epoch for Curriculum Learning	15
	GPU Used	2 Nvidia A100 with 80GB

Table 9. Hyperparameters and their values

synthetic and real data, and it also showcases lowest Out-of-Distribution (OOD) performance. As the guidance scale increases, this distribution gap diminishes, and the OOD F1 score consistently improves. This outcome aligns with the visually observed reduction in distribution differences between generated and real images.

Conversely, the trend seen with ImageNet-LT diverges from above. In long-tail classification, we aim to increase data diversity while keeping the distribution gap small. As detailed in Appendix A.2.2, on one hand, generating synthetic data that closely resemble real data further reduces the diversity, and generating synthetic data far from real distribution can offer diversity but hurt OOD performance. In case of ImageNet-LT, we observe that more diverse synthetic data tends to significantly improve the classifiers' generalization.

Inspired by these observations, we tailor our guidance scales intervals according to the task-at-hand.

A.8. Improvement on Worst-k classes: Balanced Softmax (BS) v/s DisCL with BS [29]

While DisCL's average gain over Balanced Softmax Baseline(BS) is $+2.07\%$, it improves BS's worst-k class accuracy by 4.5%–7.6%, verifying our targeted advantage on the most difficult classes—precisely where strong baselines struggle. It demonstrates that DisCL complements existing methods, improving performance where it matters most, even compared with strong baselines.

k	10	50	100	150	200
$\text{Acc}_{\text{BS+DisCL}} - \text{Acc}_{\text{BS}}$	7.6%	6.0%	5.7%	5.2%	4.5%

Table 10. **Improvement in Accuracy on Worst-k classes in INLT**.

A.9. Societal Impact

Our proposed method is beneficial for diverse fields, where inadequate quantity and low quality of data is common, *e.g.* medical domain. The synthetic data generation, as followed by DisCL approach can reduce the need for extensive data

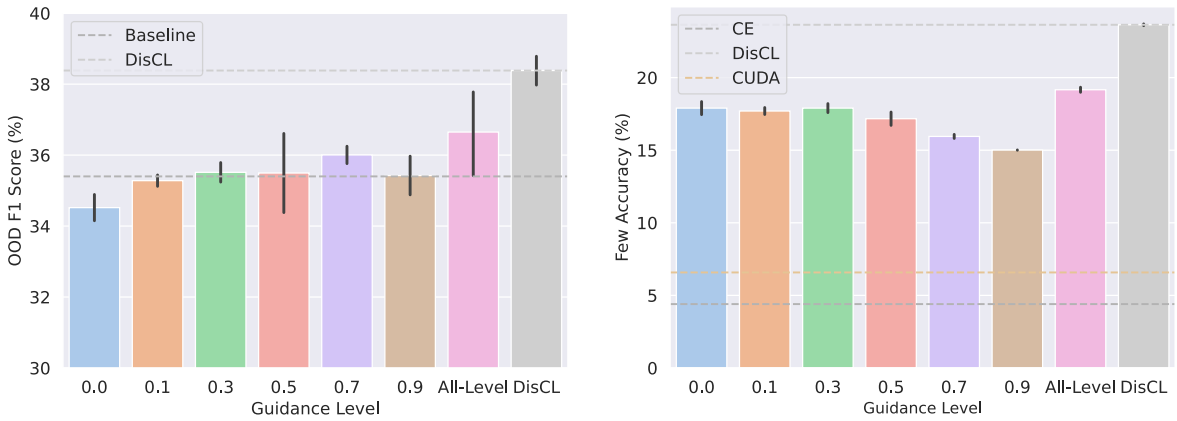


Figure 14. Effect of Image Guidance (mixing syn+real). All-level experiments use the synthesis samples from all guidance scales selected for each task. 0.5 refers to only using synthetic data with guidance level $\lambda = 0.5$ for fine-tuning. Left: results on iWildCam. Right: results on ImageNet-LT

collection, therefore mitigating the ethical concerns related to data-privacy. Overall, our method DisCL can democratize the access of effectively training ML models in the low-resource environments. However, by leveraging the pre-trained generative models, the potential biases of models can perpetuate into the synthetic data and eventually affect the sensitive real-world applications consuming this data, such as medical diagnosis, law enforcement *etc.*