Instruct Where the Model Fails: Generative Data Augmentation via Guided Self-contrastive Fine-tuning

Weijian Ma[1]*, Ruoxin Chen[2], Keyue Zhang[2], Shuang Wu[2], Shouhong Ding[2]

[1]School of Computer Science, Fudan University
[2]Youtu Lab, Tencent
mawj22@m.fudan.edu.cn

Abstract

Data augmentation is expected to bring about unseen features of training set, enhancing the model's ability to generalize in situations where data is limited. Generative image models trained on large web-crawled datasets such as LAION are known to produce images with stereotypes and imperceptible bias when used to augment training data, owing to dataset misalignment and the generator's ignorance of the downstream model. We improve downstream task awareness in generated images by proposing a task-aware fine-tuning strategy that actively detects failures of downstream task in the target model to fine-tune the generation process between epochs. The dynamic fine-tuning strategy is achieved by (1) inspecting misalignment between generated data and original data via VLM captioners and (2) adjusts both prompts and diffusion model so that the strategy dynamically guides the generator by focusing on the detected bias of VLM. This is done via re-captioning the overfitted data as well as finetuning the diffusion trajectory in a contrastive manner. To cooperate with the VLM captioner, the contrastive fine-tuning process dynamically adjusts different parts of the diffusion trajectory based on detected misalignment, thus shifting the the generated distribution away from making the downstream model overfit. Our experiments on few-shot class incremental learning show that our instruction-guided finetuning strategy consistently assists the downstream model with higher classification accuracy compared to generative data augmentation baselines such as Stable Diffusion and GPT-4o, and state-of-the-art non-generative strategies.

Introduction

The selection of practice content is a central issue in the educational process and a subject of ongoing scientific research (QIN et al. 2023; Council 2002; Meijers and Verbeek 2020). For training neural networks when data is not sufficient, data augmentation is a commonly-used technique. In this case, human use rules or generative models to generate more content, which introduces variations to training data while remaining meaningful and invariant features (LeCun et al. 1998). These methods now serve as a critical technique in machine learning, particularly when training data is scarce or hard to obtain. The aim of data augmentation is to tell the data intrinsics from unnecessary contents, preventing the model from overfitting on unrelated characteristics of training set (Baird 1995; Shorten and Khoshgoftaar 2022).

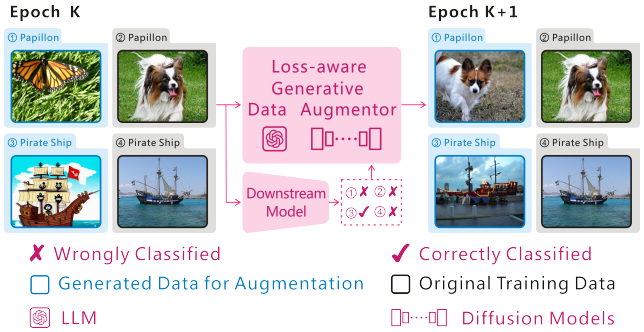

Figure 1: Method outline. Our proposed data augmentation strategy dynamically adjusts augmented images between epochs. The strategy consists of a VLM captioner and a diffusion-based image generator, which is guided by the failures of the validation set. The generative augmenter adjusts image augmentations from general semantics to inexpressible details.

Traditionally, data augmentation applies handwritten rules to transform the input data by introducing diverse variations such as rotations, translations, or noise to generate new data (LeCun et al. 1998). Such techniques force the model to focus on the consistent underlying patterns across different versions of the same data. This to some extent aligns with the principle of learning robust features, but the variation still lies in superficial stages such as colors and textures. For augmentations under different semantics, such as a cat with different poses or a person in different backgrounds, such augmentation methods still suffer.

The recent advances of generative models (Ho, Jain, and Abbeel 2020) have profoundly changed the data creation process. Generative models are now being leveraged to create synthetic data, which is particularly valuable in scenarios where labeled data is scarce. Text-to-image models, such as Stable Diffusion (Rombach et al. 2022), enables the generation of high-quality images that retain essential features of the original data while introducing controlled variations. However, the bias of diffusion models remains fixed

*Work partly done during an internship at Youtu Lab, Tencent.

and is hard to control, thus introducing imperceptible and persistent noises, hindering downstream model performance trained on such augmented data (Sarıyıldız et al. 2023).

Thinking about the human education process, the content of the practice is delicately conceived by experienced teachers familiar with student behavior. It is also dynamically adjusted and regenerated according to the learning progress of different students. But in the case of creating augmentation contents, both generative augmentation and handwritten rules remain static, and are even detached from the downstream model to be trained. Thus, the fixed augmentation strategy is actually unaware of what downstream model does not know. For example, generating too many butterflies instead of dogs for the category papillon, or rotating a curled cat for many times while makes the stretched cat unseen (Sarıyıldız et al. 2023). Hence an efficient data augmentation strategy needs to receive the feedback of downstream model and instruct where the model fails. This calls for our proposed dynamic data augmentation strategy, where the augmentation varies from general concepts to imperceptible details. The augmentation strategy also varies between different epochs, dynamically controlled by a VLM captioner.

Specifically, as is shown in Figure 1 and Figure 2, our augmentation strategy works as follows. In each epoch, data in each category are divided by correct or faulty classification. All correctly classified augmentation data and falsely classified data in the original set of each category are compared by a VLM-controlled coordinator. The coordinator then detects the deviation between the two pictures and adjusts the diffusion-based generator via a contrastive finetuning process for both diffusion trajectory and language prompts between different epochs. If the deviation lies in semantic level, such as butterflies and dogs with the same name, diffusion prompts are generated again and the first several steps of the diffusion strategies are pushed away from the falsely generated pictures. If deviation lies in imperceptible details, the latter steps of diffusion trajectory will be drawn near to correctly classified figure. The diffusion trajectory is finetuned between epochs until downstream model convergence.

Our experimental results on few-shot class incremental learning (FSCIL) demonstrate that our instruction-guided finetuning approach consistently enhances the downstream model's classification accuracy throughout the continual learning process. This improvement surpasses the performance achieved by generative data augmentation methods, including Stable Diffusion and GPT-4o, as well as state-of-the-art FSCIL strategies.

Our contribution can be summarized as follows.

- We are the first to take advantage of the performance of the downstream model as feedback during the training process, and design a dynamic data enhancement strategy that adjusts the generation process between epochs.
- VLMs are used as a coordinator of the generative augmentation process, dynamically adjusting the generation process for both text prompts and the diffusion trajectory.
- A contrastive finetuning strategy is utilized for different levels of augmentation, from semantics to details.
- The results of the experiment show that our finetuning strategy improves the classification performance of downstream models in data-scarce scenarios.

Related Work

Data Augmentation. Data Augmentation is designed to prevent the model from overfitting to unrelated characteristics of training data. It is a long-standing technique that can be traced back to (Baird 1995). Foundational techniques of data augmentation, such as rotation, scaling, translation, blurring, and noise addition, take roots in early computer vision (LeCun et al. 1998) and deep learning research and gains widespread popularity in AlexNet (Krizhevsky, Sutskever, and Hinton 2012; Shorten, Khoshgoftaar, and Furht 2019; Shorten and Khoshgoftaar 2023, 2022). Other data augmentation techniques include random erasing (Zhong et al. 2020) and histogram equalization (Pizer et al. 1987). Similarly, data augmentation in NLP domains includes synonym replacement, random insertion or deletion, random swapping, and sentence shuffling (Feng et al. 2021). However, all of these methods remain fixed during the training process. They also hardly bring variations at the semantic level.

Generative Models. Diffusion models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015) have opened up a new era of image generation after GAN series (Goodfellow et al. 2014; Arjovsky, Chintala, and Bottou 2017). They have introduced unprecedented levels of granularity in user control, allowing for highly customized image generation and editing. Stable Diffusion (Rombach et al. 2022) operates by iteratively refining an image through a series of denoising steps, enabling users to specify details at multiple levels of granularity. Recent techniques like Latent Diffusion (Rombach et al. 2022) and Iterative Multi-Granular Editing (Joseph et al. 2024) have enhanced this control further, allowing for precise, localized edits and iterative adjustments. These methods are grounded in energy-based models, which enable fine-tuning of specific regions in an image based on user-defined constraints. It is worth mentioning that works about user control focus on aligning the generation results with user intents (Liang et al. 2024) or given prompts (Kondapaneni et al. 2024). Few work ever tried to align the generation result with training effect of downstream models.

Training on Synthetic Data. Training on synthetic data has a long history where the first attempts can be traced back to (Perez and Wang 2017). It is becoming increasingly important for overcoming the challenges of inadequate real-world datasets, since it can generate labeled data economically. This technique has been widely applied to industrial anomaly detection (Wang et al. 2024) and autonomous driving (Wen et al. 2024) where labeled data is scarce. However, training on synthetic data has its drawbacks. Recent attempts (Sarıyıldız et al. 2023) tried to train ImageNet (Deng et al. 2009) classification on fully synthetic data via minimal prompt engineering. However, the downstream model fails to catch up with the performance on models trained on ImageNet itself even when far more generated data is used. This is because all these training methods use a static data generation strategy which is predefined by humans, thus neglecting the inherent domain gap between images generated

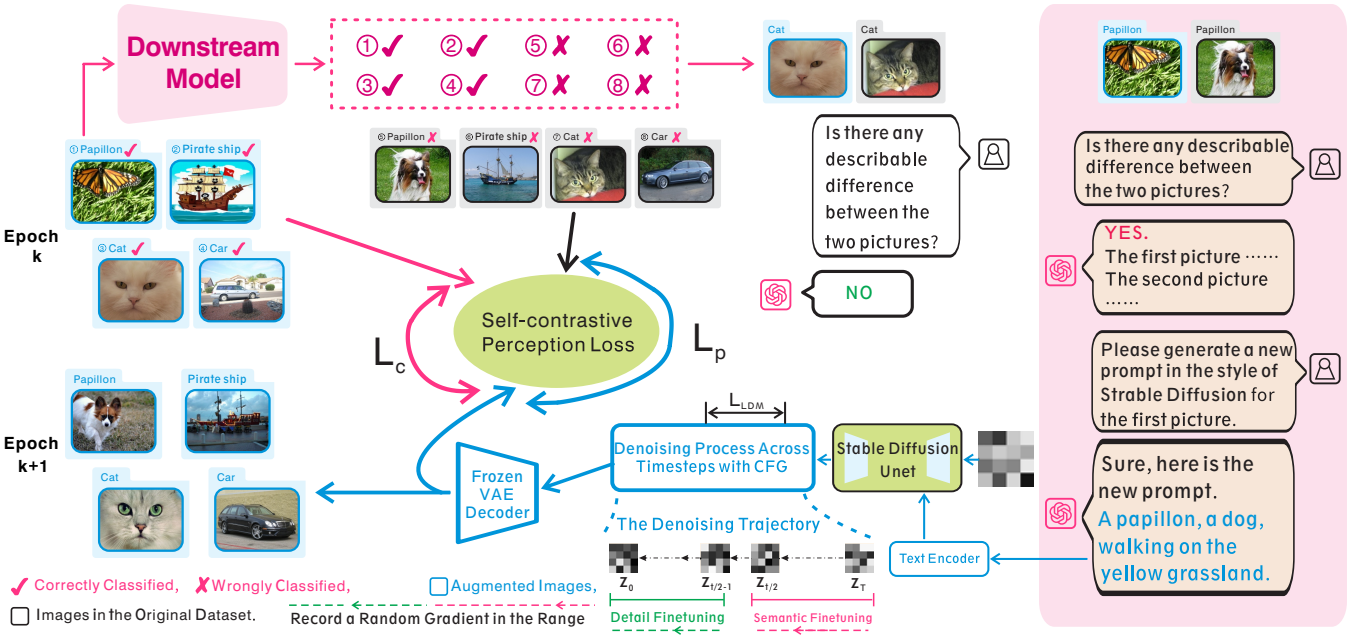

Figure 2: Method Illustration. We examine the images correctly classified in the generated augmentation set and those falsely classified in the original dataset. For an augmented image and an original image, they are first compared at semantic level to judge whether there *exists describable semantic differences* between the two. If so, the prompt of Stable Diffusion will be rewritten, and the gradient of the first several steps in the diffusion trajectory will be tuned. If no describable semantic differences exist, the prompt will remain the same, and the gradient of the last several steps will be tuned in the hope of adjusting the inexpressible details. Note that the prompts are for illustrative purposes. Detailed prompts are shown in Figure 3.

by Stable Diffusion and the original dataset, not to mention dynamically adjusting the training data to help the downstream model grasp the underlying knowledge conveyed in the training set.

Method

Method Overview

Our augmentation strategy learns from the feedback of the downstream model, and pays special attention to the classification failures on the original dataset.

In particular, at the end of an epoch k, the wrongly classified data in the validation split of *the original dataset* X_o is examined, along with *the generated augmentation which are correctly classified*, denoted as $\hat{X}_a$. X_o and $\hat{X}_a$ are the targets for contrastive finetuning of the generative model $\mathcal{G}$ which is controlled by a VLM coordinator $\Phi(\hat{X}_a, X_o)$ which detects semantic misalignment. For $\hat{x}_a$ with semantic misalignment, the prompt $c_{\hat{x}_a}$ will be rewritten and $\mathcal{G}$ will be fine-tuned semantically. $\mathcal{G}$ will be fine-tuned focusing on the image details for other samples in $\hat{X}_a$.

The procedure of our method is shown in Figure 2. The illustration of the VLM coordinator is in the next section and the fine-tuning strategy of the diffusion generator will be introduced afterwards.

The VLM Coordinator

Data generated via Stable Diffusion is known to be biased owing to the imbalance of web dataset like LAION (Schuhmann et al. 2022). Therefore, a checking strategy is required when diffusion generated data is used in training downstream models to avoid semantic bias. We adopt prevailing large VLMs as the coordinator of the diffusion generator, which checks the potential semantic misalignment between generated data, modifies the prompt for semantic misalignment, and selects the generation stage of the diffuser to be finetuned via choosing a specific timestep.

To avoid the cost of comparing $\hat{X}_a$ and X_o one by one, correctly classified augmentation data is sampled via farthest point sampling and clustering (Qi et al. 2017) method, with similarity measurement as negative SSIM (Wang et al. 2004). A typical checking strategy for misalignment is shown in Figure 3. In epoch k, for each selected image pair $(\hat{x}_a^k, x_o^k)$, their detailed captions $(c_{\hat{x}_a}^k, c_{x_o}^k)$ are generated via VLM captioner and are semantically compared. If the two captions are not semantically similar, the VLM is asked to generate the new prompt $c_{\hat{x}_a}^{k+1}$. Formally, the checking strategy is defined as

$$c_{\hat{x}_a}^{k+1} = \begin{cases} \mathcal{P}(c_{\hat{x}_a}^k, c_{x_o}^k) & \text{if } \sum_{x_o \in X_o} \Phi(\hat{x}_a, x_o) > 0 \\ c_{\hat{x}_a}^k & \text{otherwise} \end{cases} \tag{1}$$

$$\Phi(\hat{x}_a, x_o) = \begin{cases} 0 & \text{if } \hat{x}_a \text{ and } x_o \text{ are semantically similar.} \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

Please describe the following image with special attention to category, style, and pose and background.

The image depicts a green butterfly, which appears to be a painted or digitally rendered illustration rather than a photograph.

Category:

.....

Overall, the image has a serene and elegant quality, likely intended to highlight the beauty and delicate nature of the butterfly.

Please describe the following image with special attention to category, style, and pose and background.

The image shows a small dog sitting on a grassy lawn.

Category:

.....

Overall, the image captures a warm and lively moment, with the dog appearing both relaxed and attentive in a natural environment.

Are the two images semantically similar? Please answer YES or NO. Do not output extra contents.

NO

You are a prompt engineer of Stable Diffusion, the corresponding prompt of the first image is "A papillion". Please modify the prompt of the first image so that Stable Diffusion can generate the second image. Add necessary contents if needed.

"A Papillon dog sitting on a grassy lawn, facing the camera with its long, flowing fur and distinctive black-tipped ears. The dog has a reddish-brown and white coat, with a friendly expression, captured in a realistic outdoor setting with a soft, natural background."

Figure 3: Illustration of the proposed VLM Coordinator checking semantic misalignment of the augmented image based on the promopt '[A papillon]'. Details of the process are listed in Supplementary Materials for space limit.

where $\mathcal{P}$ is the prompt modifier. and Φ is achieved via the prompting process in Figure 3.

Guided by the newly generated prompt, generation result of $\mathcal{G}$ will semantically approach the failures in the original dataset, namely where the model fails in. For coordinating the finetuning process of the downstream diffusion model $\mathcal{G}$, each image with modified prompt will be labelled as semantic fine-tuning, and detail fine-tuning will be labelled for the remaining images, as illustrated below.

The Contrastive Finetuning Strategy

With the label of semantic finetuning or detail finetuning generated by the VLM coordinator, the objective is to tune Stable Diffusion for generating images that is wrongly classified by downstream model. The finetuning strategy is based on Low-Rank Adaptation (LoRA) (Hu et al. 2021) to prevent overfitting on small training data and lower the memory demands since only parameter of new low-rank weight matrices are trained.

Below we will first briefly introduce latent diffusion models, and how to measure the perceptual misalignment between the generated augmentation and the target image, namely the wrongly-classified ones. And finally the self-contrastive loss of the fine-tuning strategy will be introduced.

Latent Diffusion Model. Diffusion models iteratively reverses a forward noising process and fits the outcome into a desired data distribution. Latent diffusion models conducts the generation process in the latent space of data. For each timestep t, the latent noise is defined as $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}$, where $\mathbf{z}_0$ is the latent variable encoded from real image and α_t controls the strength of gaussian noise ϵ. We make use of the pretrained denoising network $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)$, which predicts the partially denoised latent to obtain $\mathbf{z}_{t-\delta t}$ under text prompt $\mathbf{c}$. The training objective of the denoising network ϵ_θ is to minimize the predicted noise:

$$\mathcal{L}_{\text{LDM}}(\mathbf{z}, \mathbf{c}) = \mathbb{E}_{\epsilon, \mathbf{z}, \mathbf{c}, t}\left[w_t \| \epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) \|\right] \tag{3}$$

where w_t controls the noise schedule.

For conditional generation based on text prompts, we adopt classifier-free guidance (Ho and Salimans 2022), which iteratively denoises the latent code as follows

$$z_{t-1} = (1 + w(t))\, \epsilon(\mathbf{z}_t, \mathbf{c}_k, t) - w(t)\epsilon(\mathbf{x}_t, \emptyset, t), \tag{4}$$

where $w(t)$ is the guidance scale, $\mathbf{c}_k$ is the prompt of the augmented image at epoch k and $\emptyset$ is a null text embedding.

Preceptural Misalignment Measurement. $\mathcal{L}_{\text{LDM}}$ operates on the latent codes of diffusion models. Such manipulation is desirable for adding expressible concepts such as adding an object, or changing the general style of the image. This can be applied to semantic finetuning of training data. However, such general control does not apply to intricate concepts that is not expressible by captions. In this sense, we decode the latent variable $\mathbf{z}_0$ to image space and propose to use an image perceptual loss $\mathcal{L}_P$ which focuses on the luminance, contrast, and structure of the generated augmentation $\hat{x}$ and the target image x, which is defined as

$$\mathcal{L}_P(\hat{x}, x) = \mathbb{E}_{\hat{x}, x}\left[\text{SSIM}(\hat{x}, x)\right] \tag{5}$$

where SSIM refers to the Structural Similarity Index.

It is not trivial to backpropagate the perceptual loss in image space to the iterative denoising steps of the latent diffusion model. Using $\mathcal{L}_P$ to finetune the entire decoder and diffusion model not only requires significant memory and training time, but also results in network instability when finetuning data is scarce. To address this, we alter the gradients based on different finetuning schemes decided by the VLM coordinator. For semantic finetuning, we randomly record the gradient of one of the first half of the gradient steps. While for detail finetuning, we record a random gradient step at the latter half of the gradients. The fine-tuning is conducted on the exacted diffusion step recorded, and the gradients on the other steps remain frozen. Our findings reveal that gradients from the first several steps influences the semantics of generated images most, while the later steps manipulates the imperceptible details. This aligns with the explanations of (Namekata et al. 2024) and (Guo et al. 2024).

Self-contrastive finetuning strategy The perceptural loss illustrates the fine-tuning directions for Diffusion models to align with the positive examples. To amplify the gradient adjustment effect of the perceptural loss, we design a contrastive finetuning strategy of Stable Diffusion based on the following insights.

- The correctly classified augmented examples in epoch k are natural negative examples when augmenting data for epoch $k+1$ since overfitting on augmented data is not desirable.
- The wrongly classified data in the original dataset in epoch k are what the downstream model needs to learn in epoch $k+1$. The augmentation at epoch $k+1$ should approach samples where downstream model stucks at.

One thing needs to consider is that the correctly classified augmented examples is far more than the wrongly classified original data. In this sense, we utilize a straightforward sampling strategy for hard negatives, thereby emphasizing that the primary distinctions lie in the diffusion model's misalignment between generated data and downstream training data. We achieve this by using a simple selection rule based on SSIM. In particular, the augmented images with the largest SSIM between the wrongly classified original data is selected as the negative examples. In practice, 8 negative examples are selected for each positive example.

To enhance the error-based fine-tuning process, our objective is to ensure that images generated by model at epoch $k+1$, denoted as $\hat{x}^0_{\theta_{k+1}}$ have closer perceptual distances to positive examples and farther distances from negative examples. This reinforces the model's alignment with wrongly classified examples in the original dataset, distancing itself from the overfitted features in the downstream models. This can be formally expressed as

$$\text{SSIM}(\hat{x}^0_{\theta_{k+1}}, x^+) > \text{SSIM}(\hat{x}^0_{\theta_{k+1}}, x^-). \tag{6}$$

In this sense, the training objective is formulated based on triplet loss, which we name Self-Contrastive Perceptual Loss, is formally written as follows,

$$\mathcal{L}_C(\hat{x}, x^+, x^-) = \mathbb{E}_{\hat{x},x^+,x^-}[\max(\text{SSIM}(\hat{x}^0_{\theta_{k+1}}, x^+) - \lambda\text{SSIM}(\hat{x}^0_{\theta_{k+1}}, x^-) + m, 0)] \tag{7}$$

where λ denotes the weights on negative examples, m is the constant margin between positive and negative examples,

In all, the final loss of finetuning Stable Diffusion is the weighted sum of $\mathcal{L}_{\text{LDM}}$, $\mathcal{L}_P$ and $\mathcal{L}_C$

$$\mathcal{L}_\Theta = \mathcal{L}_{\text{LDM}} + \beta_1\mathcal{L}_P + \beta_2\mathcal{L}_C \tag{8}$$

where β_1 and β_2 are hyperparameters.

Experiment

To justify the design choice of our augmentation strategy with diffusion models and VLM combined, we conduct experiments on few-shot class incremental learning (FSCIL) (Tao et al. 2020). This setting greatly simulates our discussed scenario where the training data is scarce and the finetuning strategy is dynamic.

Experiment Settings

Dataset and Evaluation Metrics. We conduct our experiment under the setting of (Tao et al. 2020) and (Park, Song, and Park 2024) for fair comparison. The method is evaluated with state-of-the-art method on following datasets: miniImageNet (Ravi and Larochelle 2017), CUB200 (Wah et al. 2011) and CIFAR-100 (Krizhevsky 2009). The split configurations in all datasets are shown in Table 2 which remains the same as the prevailing settings in (Tao et al. 2020) and (Park, Song, and Park 2024). 5 simulations are conducted for each experiment and the averages are reported. The performance of different methods is evaluated by measuring base session accuracy A_{Base}, last session accuracy A_{Last}, and the average accuracy across all the sessions A_{Avg}.

Implementation Details. We use Stable Diffusion v1.5 as our diffusion augmentor with the CFG guidance scale as 2, following the configuration of (Sarıyıldız et al. 2023). The total diffusion steps is set to 20. The VLM we utilize is GPT-4o in the main experiment. The initial prompt for SD 1.5 remains fixed as *A picture of a [category].*

For downstream models, we used a VIT-B/16 (Dosovitskiy et al. 2021) pretrained on ImageNet-21K (Deng et al. 2009) for ours and other comparative methods. The learning rate of the downstream model is set as $2e-4$, using the Adam optimizer and cosine annealing learning rate scheduler. For each image in the training set, we augment the number of images to the original size in each epoch for augmentation. For both base session and incremental sessions, we train our network until convergence. The method is trained on 8 H100-80G GPUs.

Comparative methods. We set finetuning the backbone with a new classification head as a baseline method. We also include some following recent SOTA FSCIL methods for comparison, including CEC (Zhang et al. 2021), WaRP (Kim et al. 2023), NC-FSCIL (Yang et al. 2022), L2P (Wang et al. 2022b), DualPrompt (Wang et al. 2022a), and PriViLege (Park, Song, and Park 2024). L2P, DualPrompt and

Dataset	CUB200			CIFAR-100			miniImageNet		
Method	A_{Base}	A_{Last}	A_{Avg}	A_{Base}	A_{Last}	A_{Avg}	A_{Base}	A_{Last}	A_{Avg}
Fine-Tuning + Proto ψ	84.21±0.13	3.79±1.47	21.60±1.32	91.36±0.15	5.19±0.13	37.04±1.06	93.67±0.02	9.87±5.42	44.60±0.92
CEC [CVPR'21]	75.40±8.01	65.70±8.03	72.41±1.18	74.20±2.03	61.48±3.33	67.10±2.92	87.43±5.90	80.74±7.51	83.06±7.14
L2P [CVPR'22]	44.97±2.32	15.41±3.45	24.99±4.30	83.29±0.50	49.87±0.31	64.08±0.39	94.59±0.21	56.84±0.32	72.97±0.36
DualPrompt [ECCV'22]	53.37±1.83	23.25±2.02	36.30±2.39	85.11±0.29	50.93±0.21	65.45±0.27	95.05±0.20	57.14±0.11	73.31±0.15
NC-FSCIL [ICLR'23]	78.49±2.32	38.80±1.14	57.92±1.71	89.51±0.23	53.70±0.14	68.96±0.17	77.25±0.42	46.35±0.25	59.52±0.33
WaRP [ICLR'23]	67.74±5.57	49.36±6.56	55.85±6.06	86.20±1.46	65.48±1.87	74.55±1.67	83.30±1.06	67.97±1.28	74.13±1.08
PriViLege [CVPR'24]	82.21±0.35	**75.08±0.52**	77.50±0.33	90.88±0.20	86.06±0.32	88.08±0.20	96.68±0.06	94.10±0.13	95.27±0.11
Ours	**86.73±0.67**	73.82±0.74	**79.63±0.68**	**95.57±0.23**	**87.86±0.45**	**90.01±1.01**	**97.35±0.24**	**95.23±0.33**	**96.54±0.41**

Table 1: Comparison of the performance on CUB200, CIFAR-100, and miniImageNet. CUB200 has a 10-way 5-shot incremental setup, and CIFAR-100 and miniImageNet have a 5-way 5-shot incremental setup. We report the best as **bold** and the second-best as underlined.

Session	CUB200	CIFAR-100	miniImageNet
Base	100	60	60
Incremental	10-way 5-shot	5-way 5-shot	5-way 5-shot
# of sessions	1+10	1+8	1+8

Table 2: Configuration settings for FSCIL benchmarks on CUB-200, CIFAR-100, and miniImageNet.

PriViLege use ViT for classification while the other three use some methods else. The comparisons against previous SOTAs are included in the main experiment. We also provide some baselines where only finetuning SD, or using GPT-4o for zero-shot classification. These comparisons will be put forward in ablation study.

Main Results

The best, the last, and the average accuracy of CUB200, CIFAR-100, and miniImageNet are reported in Table 1 respectively. From Table 1 we can observe that, our method, as a method based on finetuning, not only overwhelmed most of the baselines by a visible margin on all the benchmarks compared with previous SOTA methods in the field of FSCIL. Our proposed method, assisted by the strong perception ability of large VLMs as well as the fine-grained fientuning strategy of diffusion models, has reported a $\sim 3\%$ performance gain on all benchmarks despite the well-performing baseline achieving a high base performance of at most 94%. This showcases that a fine-grained curriculum based on training data selection can reveal the obtained knowledge of the pretrained model to the fullest, also making it grasp new knowledge of the unseen features on new images in easier ways. This also showcases that knowledge conveyed in these datasets can be grasped by VLM and conveyed in language which can be understood by the diffusion generator.

It is also worth mentioning that the terrible finetuning performance of the last, and the average accuracy have been saved by our finetuning strategy. This is because of the following reasons.

- First, the FSCIL testing scheme is a challenging scheme for finetuning, especially when the incremental training data is too scarce to feed the classification head, leading to the divergence of downstream model. This illustrates the importance of using data augmentation properly in real-world scenarios when training data is scarce.
- On the other hand, what we should do is introduce real semantics to training set instead of merely cropping or jittering the image in a human-defined manner and fill up the entire training set with low-quality duplicates.

However, our proposed augmentation strategy still suffers a little at CUB200 dataset. This is because the semantics of birds lies rather at the detail level, and the caption of VLM has perturbed the performance of diffusion generator a little bit. Our generator tries to perceive the high-level semantic difference between different birds, instead of merely digging the statistical differences in the dataset.

Visualization of the Augmentation Process

In this part we put forward an example visualization of the augmented training process. We exhibit the augmentation process from the viewpoint of how the generated image of a certain category, as well as its corresponding prompt, changes between epochs. The dynamic of the augmentation is shown in Figure 4. From the dynamics we obtain the following observations.

- The prompts remain fixed after the first several epochs, which showcases that the semantic misalignment between the diffusion model and the training dataset can be merged quickly thanks to the VLM captioner. This also avoids the problem posted by (Sarıyıldız et al. 2023).
- There are still visible differences between the generated images at different epochs when the prompt remains fixed. This showcases that our diffusion model brings about substantial diversity to the training set, thus bringing more diversity to downstream models.

In all, the dynamics in Figure 4 shows that our generative augmentation strategy can add diversity from semantics to

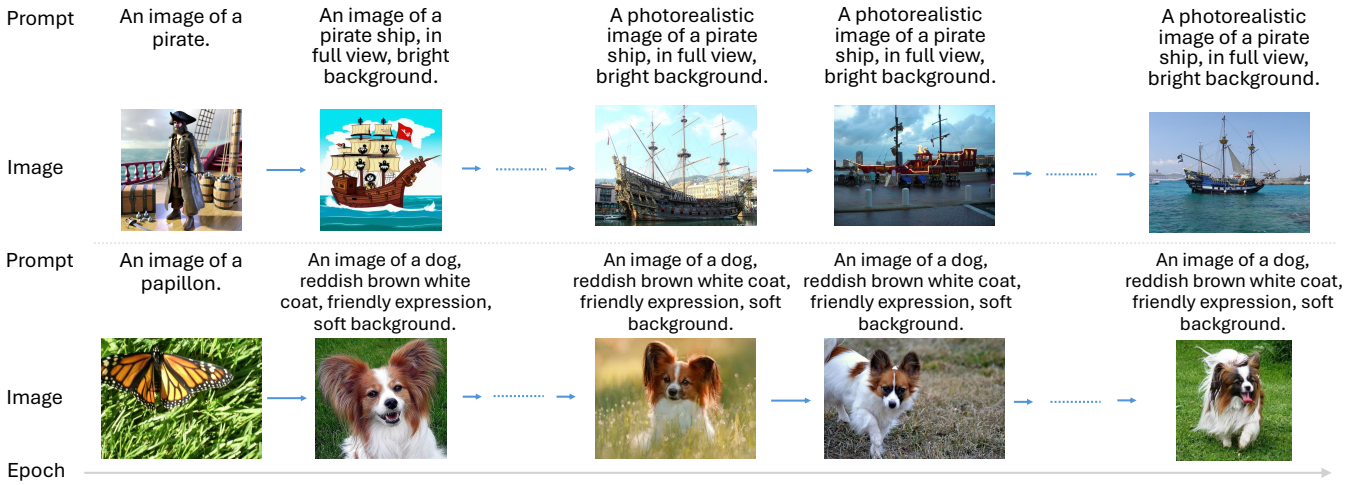

Figure 4: Visualization of the dynamics of our proposed augmentation strategy. We sample prompts and the corresponding images from different classes and illustrate its variation between different epochs. The dynamics show that the change of the prompt, as well as the general semantics, happen at the first few epochs. The details of the images still vary slightly at the latter epochs, bringing variations to the downstream training process.

Dataset	miniImageNet		
Ablation	A_{Base}	A_{Last}	A_{Avg}
Baseline	93.67±0.02	9.87±5.42	44.60±0.92
+ Fixed SD	94.10±0.87	48.37±3.76	64.26±0.77
+ Random	94.33±1.37	50.17±4.25	64.18±1.48
+ First	94.43±0.92	50.67±2.83	66.36±0.62
+ VLM	96.98±0.86	94.87±0.71	96.31±0.60
Ours	**97.35±0.24**	**95.23±0.33**	**96.54±0.41**

Table 3: Ablation experiment on miniImageNet. The baseline denotes fine-tuning pre-trained ViT with prototype classifier ψ. + denotes adding a component from the previous line. Random refers to finetuning a random gradient in diffusion trajectory. Semantic refers to finetuning first gradients only. VLM refers to re-captioning via VLM coordinator.

detail, thus bringing diversity to downstream models while avoiding misalignments.

Ablation Study

We perform two ablation studies on miniImagenet dataset and CUB200 dataset to justify our design of finetuning on both semantic level and in details.

We first justify the design of finetuning on semantic level. We conduct this ablation on the miniImagenet dataset. The result is shown in 3 From the last accuracy we can observe that even fixed SD with bias does bring about substantial improvements to the model since it makes the network converge. However, tuning SD with no language guidance does not improve much since the semantic gap between a dog and a butterfly is too big to merge via contrastive learning. The semantic gap across labels should be merged via directly changing prompts, as adding the VLM shows.

For the justification of detail finetuning, we can partially observe the result from using the VLM alone to conduct the classification task. It can be seen that VLM is not capable

Dataset	CUB 200		
Ablation	A_{Base}	A_{Last}	A_{Avg}
VLM	56.78±0.01	56.73±0.03	56.75±0.02
Baseline	84.21±0.13	3.79±1.47	21.60±1.32
+SD+VLM	84.75±0.25	54.86±2.38	60.47±3.02
Ours	**86.73±0.67**	**73.82±0.74**	**79.63±0.68**

Table 4: Ablation experiment on CUB 200. The baseline and + remain the same as above. SD+VLM refers to adding a fixed SD with dynamic VLM prompts.

of handling such intricate differences between birds since it may not give any option. This also makes augmenting with SD and the VLM prompt vague, illustrating the importance of detailed finetuning.

Conclusion

This work draws inspiration from the human learning process and redesigns the data augmentation strategy where the feedback of downstream model is taken into account throughout the generation process. In particular, a VLM-based coordinator is designed to detect the semantic misalignment between the augmented images and images in the original dataset. The generation prompt is re-generated when semantic misalignment exists. A contrastive finetuning strategy is proposed which measures the perceptual misalignment between generated images and wrongly classified target image, and guides the generation trajectory towards the target image via a contrastive learning process. The results show that our augmentation strategy can achieve better performance in few-shot scenarios while using fewer generated data. This illustrates the importance of the fine-grained perception for the feedbacks of downstream models, as well as the effectiveness of our model design. Future work includes extending our paradigm into more scenarios such as detection and segmentation.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

Baird, H. S. 1995. Document Image Defect Models. In *Document Image Analysis*, 315–325. Los Alamitos, CA, USA: IEEE Computer Society Press.

Council, N. R. 2002. *Scientific Research in Education*. Washington, DC: The National Academies Press.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A Survey of Data Augmentation Approaches for NLP. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988. Online: Association for Computational Linguistics.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, 2672–2680.

Guo, L.; Chen, W.; Sun, Y.; Ai, B.; Pappas, N.; and Quek, T. 2024. Diffusion-Driven Semantic Communication for Generative Models with Bandwidth Constraints. *arXiv preprint arXiv:2407.18468*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2006.11239*.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Wang, L. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Joseph, K. J.; Udhayanan, P.; Shukla, T.; Agarwal, A.; Karanam, S.; Goswami, K.; and Srinivasan, B. V. 2024. Iterative Multi-Granular Image Editing using Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Kim, D.-Y.; Han, D.-J.; Seo, J.; and Moon, J. 2023. Warping the space: Weight space rotation for class-incremental few-shot learning. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Kondapaneni, N.; Marks, M.; Knott, M.; Guimaraes, R.; and Perona, P. 2024. Text-Image Alignment for Diffusion-Based Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13883–13893.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, 1097–1105.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; Ke, J.; Dvijotham, K.; Collins, K.; Luo, Y.; Li, Y.; Kohlhoff, K. J.; Ramachandran, D.; and Navalpakkam, V. 2024. Rich Human Feedback for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Meijers, A.; and Verbeek, P.-P. 2020. Epistemological and educational issues in teaching practice-oriented scientific research: roles for philosophers of science. *European Journal for Philosophy of Science*, 10(3): 42–58.

Namekata, K.; Sabour, A.; Fidler, S.; and Kim, S. W. 2024. EmerDiff: Emerging Pixel-level Semantic Knowledge in Diffusion Models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Park, K.-H.; Song, K.; and Park, G.-M. 2024. Pre-trained Vision and Language Transformers Are Few-Shot Incremental Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23881–23890.

Perez, L.; and Wang, J. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *CoRR*, abs/1712.04621.

Pizer, S. M.; Amburn, E. P.; Austin, J. D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B. T.; Zimmerman, J. B.; and Zuiderveld, K. 1987. Adaptive Histogram Equalization and Its Variations. In *Computer Vision, Graphics, and Image Processing*, volume 39, 355–368.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, 5099–5108.

QIN, H.; YAN, B.; LIN, S.; ZHOU, L.; and XIAO, Z. 2023. Reform and Practice of the Practical Course about Scientific Research for Undergraduates. *Experiment Science and Technology*, 21(6): 99–105.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *International conference on learning representations (ICLR)*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*.

Sarıyıldız, M. B.; Alahari, K.; Larlus, D.; and Kalantidis, Y. 2023. Fake It Till You Make It: Learning Transferable Representations from Synthetic ImageNet Clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10710–10720.

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. *arXiv preprint arXiv:2210.08402*.

Shorten, C.; and Khoshgoftaar, T. M. 2022. Image Data Augmentation for Deep Learning: A Survey. *arXiv preprint arXiv:2204.08610*. Available at: https://arxiv.org/abs/2204.08610.

Shorten, C.; and Khoshgoftaar, T. M. 2023. Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions. *arXiv preprint arXiv:2301.02830*. Available at: https://arxiv.org/abs/2301.02830.

Shorten, C.; Khoshgoftaar, T. M.; and Furht, B. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1): 1–48.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Bach, F.; and Blei, D., eds., *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2256–2265. Lille, France: PMLR.

Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; and Gong, Y. 2020. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, C.; Zhu, W.; Gao, B.-B.; Gan, Z.; Zhang, J.; Gu, Z.; Qian, S.; Chen, M.; and Ma, L. 2024. Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22883–22892.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022a. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision (ECCV)*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to Prompt for Continual Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; and Zhang, X. 2024. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6902–6912.

Yang, Y.; Yuan, H.; Li, X.; Lin, Z.; Torr, P.; and Tao, D. 2022. Neural Collapse Inspired Feature-Classifier Alignment for Few-Shot Class-Incremental Learning. In *The Eleventh International Conference on Learning Representations*.

Zhang, C.; Song, N.; Lin, G.; Zheng, Y.; Pan, P.; and Xu, Y. 2021. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13001–13008.