Diffusion-Classifier Synergy: Reward-Aligned Learning via Mutual Boosting Loop for FSCIL

Ruitao Wu[1,2] **Yifan Zhao**[1]* **Guangyao Chen**[3] **Jia Li**[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE & QRI, Beihang University
[2]Zhongguancun Academy [3]Peking University
{ruitaowu, zhaoyf}@buaa.edu.cn

Abstract

Few-Shot Class-Incremental Learning (FSCIL) challenges models to sequentially learn new classes from minimal examples without forgetting prior knowledge, a task complicated by the stability-plasticity dilemma and data scarcity. Current FSCIL methods often struggle with generalization due to their reliance on limited datasets. While diffusion models offer a path for data augmentation, their direct application can lead to semantic misalignment or ineffective guidance. This paper introduces Diffusion-Classifier Synergy (DCS), a novel framework that establishes a mutual boosting loop between diffusion model and FSCIL classifier. DCS utilizes a reward-aligned learning strategy, where a dynamic, multi-faceted reward function derived from the classifier's state directs the diffusion model. This reward system operates at two levels: the feature level ensures semantic coherence and diversity using prototype-anchored maximum mean discrepancy and dimension-wise variance matching, while the logits level promotes exploratory image generation and enhances inter-class discriminability through confidence recalibration and cross-session confusion-aware mechanisms. This co-evolutionary process, where generated images refine the classifier and an improved classifier state yields better reward signals, demonstrably achieves state-of-the-art performance on FSCIL benchmarks, significantly enhancing both knowledge retention and new class learning.

1 Introduction

Class-Incremental Learning (CIL) [65, 74, 58, 81, 31, 78] endeavors to equip models with the ability to learn new classes sequentially without forgetting previously acquired knowledge, a critical capability for real-world, dynamic environments. Few-Shot Class-Incremental Learning (FSCIL) [53, 52, 77, 50, 76] intensifies this challenge by further constraining that new classes are introduced with only a handful of training examples. This exacerbates the notorious *stability-plasticity dilemma* [34] and introduces severe *unreliable empirical risk minimization* [61] for novel classes due to data scarcity. The core of these issues lies the model's restricted access to past data and the limited information available for new concepts [32].

A common thread among mainstream FSCIL approaches is their reliance on the limited knowledge encapsulated within the initially provided datasets. This inherent limitation hinders their ability to significantly enhance both *intra-task generalization* (robustness within a learned class) and *inter-task generalization* (adaptability and discrimination across incrementally learned classes). The advent of powerful generative models, especially diffusion models [10, 16, 44, 48], offers a promising avenue to overcome these data limitations by synthesizing additional training samples. By generating images for old classes, they can facilitate knowledge replay with-

*Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-2153

out explicit storage, and by augmenting new, few-shot classes, they can provide richer training signals. The strong generalization capabilities of pre-trained diffusion models suggest they can introduce diverse and novel variations, potentially improving classifier robustness. However, the direct application of diffusion models for data generation in FSCIL is not without significant hurdles. As identified in our analysis (detailed in Section 4.1), two primary problems emerge: (i) **Semantic Misalignment and Diversity Deficiency of Generated Images** (Figure 1(a) ①②). When conditioned solely on class names, vanilla diffusion models tend to generate images with significant semantic deviations or insufficient diversity from the FSCIL dataset. This results in misrepresentative or overly concentrated distributions, which can introduce noise and distort learned decision boundaries, thereby degrading classifier performance. (ii) **Inefficient Feedback for Guiding Image Generation** (Figure 1(a) ③). Existing generative methods [2, 47] typically lack a mechanism for the diffusion model to adapt its output based on the classifier's current state or learning needs. The generation process is often *blind* to whether the synthesized samples are genuinely beneficial, too easy, or too difficult for the classifier, or whether they address critical areas of confusion in the feature space. While million-scale image generation performs well [4], its application is impractical in resource-demanding settings like FSCIL. We therefore prioritize the efficiency associated with generating fewer (< 50) images per class.

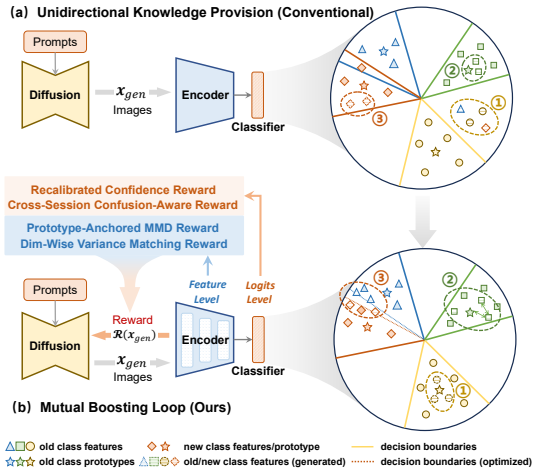

Figure 1: **(a)** Unidirectional knowledge provision in conventional methods results in inefficient generated images. **(b)** Our approach mitigates this inefficiency via a combined feature-level and logits-level reward, facilitated by a mutual boosting loop between the diffusion model and classifier.

To address these challenges, we introduce **Diffusion-Classifier Synergy (DCS)**, a novel framework that establishes a mutual boosting loop between diffusion model and FSCIL classifier. The core idea is to leverage the Diffusion Alignment as Sampling (DAS) [21] algorithm guided by a carefully designed, dynamic reward function. This reward, derived from the classifier's state, steers the diffusion model to generate strategically beneficial images. These generated images, in turn, enhance the classifier's training, leading to more refined reward signals, thus creating a co-evolutionary process.

Our DCS framework tackles the aforementioned issues at two distinct levels (Figure 1(b)): (i) At the **feature level**, where rewards are computed from features extracted by the encoder from images generated by the diffusion model, we introduce a reward design for semantic coherence and diversity (Section 4.2). This mechanism introduces two key rewards. The prototype-anchored maximum mean discrepancy reward function using MMD to encourage generated image diversity while maintaining consistency with class prototypes. Complementing this, the dimension-wise variance matching reward operates by aligning per-dimension feature variances of generated images with those from limited real data, offering a robust approach to match feature spread for new classes. This ensures that generated images are not only semantically anchored to the target class representations but also exhibit rich intra-class variations. (ii) At the **logits level**, where rewards are computed from classifier outputs for images generated by the diffusion model, we introduce a reward design for classifier-aware generation (Section 4.3). This design incorporates a recalibrated confidence reward to encourage the generation of more exploratory and generalized intra-class images, complementing the feature-level diversity reward. Building upon this, a novel cross-session confusion-aware reward is proposed, wherein the core idea is to intentionally generate *hard* samples that target the classifier's weaknesses. This is accomplished by adjusting the weight of classes in the cross-entropy based on the severity of their confusion, with the specific goal of enhancing discriminability between the target new class and its most confusable old classes.

Our contributions can be summarized as follows:

- We propose Diffusion-Classifier Synergy (DCS), a novel FSCIL framework that pioneers a mutual boosting loop between the diffusion model and the classifier, leveraging reward-aligned generation via DAS for synergistic co-evolution.

https://doi.org/10.52202/085713-2153

- We design a multi-faceted reward function operating at two distinct levels: at the feature level, it enforces semantic coherence and promotes intra-class diversity; and at the logits level, it guides the generation of exploratory, generalized intra-class images and enhances inter-class discriminability
- We empirically demonstrate state-of-the-art performance on challenging FSCIL benchmarks, showcasing significant improvements in both preserving knowledge of old classes and effectively learning new classes with limited data.

2 Related Works

Class Incremental Learning. In class-incremental learning tasks, models are required to continually learn to recognize new classes from a sequential data stream while retaining previously learned knowledge. Data replay-based methods [5, 23, 75, 59, 82] achieve this by storing data from previous tasks (such as image examples or features) or generating images of previously learned classes, allowing the model to revisit past data distributions. Network expansion-based methods [71, 64, 29, 65, 58, 79, 50] dynamically adjust the model's architecture or capacity during training to enhance its ability to learn new knowledge. Parameter regularization-based methods [69, 70, 25] focus on how the model parameters should dynamically adapt when the network structure remains fixed.

Few-Shot Class-Incremental Learning. FSCIL aims to achieve continual learning of new knowledge in data-constrained scenarios. Representative works such as [52, 73, 66, 67, 20, 38] focus on dynamic network-centric approaches, maintaining the topological relationships between feature spaces of various categories by dynamically adjusting the network structure. Methods like LIMIT [80, 8, 43] introduce the concept of meta-learning into FSCIL. Feature space-based methods [77, 41, 68, 51, 3, 27, 39, 37] enhance the robustness of the learned feature space by introducing virtual class instances and other means. Recently, many methods [40, 76, 17, 24] have employed pre-trained vision-language models (*e.g.*, CLIP) to further enhance the generalization capability of the models.

Diffusion Models for Image Classification. Data augmentation using diffusion model (DM) is an active research area aimed at enhancing image classifier training by synthesizing additional data. Key efforts focus on generating images that are both semantically consistent with class labels and exhibit sufficient richness and diversity to improve classifier performance. Initial studies [4, 35, 28, 13] demonstrated the viability of using fine-tuned DM for large-scale classification improvements, while subsequent research has explored advanced conditioning mechanisms [54] and image editing techniques [19, 62] to enhance semantic control and sample variety. Other approaches investigate leveraging DM features for precise guidance [30] or tailoring generation for specific scenarios like few-shot learning [56] or continual learning [46]. Distinct from the aforementioned approaches, our work focuses on exploring methodologies that dynamically adjust output content based on classifier feedback, with a particular emphasis on enhancing the efficiency of image generation in scenarios characterized by smaller generation scales (*e.g.*, tens of images rather than the conventional millions).

3 Preliminary

Few-Shot Class-Incremental Learning FSCIL involves sequentially learning from a continuous data stream $\mathcal{D}_{train} = \{\mathcal{D}^t_{train}\}^T_{t=0}$. Each session t introduces training samples (x_i, y_i) for a new set of disjoint classes $\mathcal{C}^t$. Crucially, after training on $\mathcal{D}^t_{train}$, the model is evaluated on all classes encountered thus far: $\mathcal{C}^t_{seen} = \bigcup^t_{s=0} \mathcal{C}^s$. FSCIL is characterized by an initial base session ($t = 0$) with ample data, followed by incremental sessions ($t > 0$) where new classes are introduced with only a few examples, typically in an N-way K-shot format.

Mainstream FSCIL methods tend to adopt an *incremental-frozen* strategy. An initial model $\sigma(x) = W^T f(x)$ (comprising a feature extractor f and classifier W) is trained extensively on base session data using a classification loss, *e.g.*, cross-entropy:

$$\mathcal{L}_{cls}\left(\sigma; x, y\right) = \mathcal{L}_{ce}\left(\sigma\left(x\right), y\right). \tag{1}$$

Subsequently, the feature extractor f is frozen. In incremental session s, the classifier weights $\boldsymbol{w}^s_c$ for new classes are typically computed as prototypes, which are defined as the average feature embeddings of their K training samples, such that $\boldsymbol{w}^s_c = \frac{1}{K}\sum^K_{i=1} f\left(x_{c,i}\right)$. The full classifier W_{full}

https://doi.org/10.52202/085713-2153

then encompasses weights for all seen classes. Inference at each session employs the Nearest Class Mean (NCM) algorithm [33]. A sample x is classified by finding the class prototype most similar to its feature embedding $f(x)$: $c_x = \arg\max_{c,s} \text{sim}\left(f(x), \boldsymbol{w}_c^s\right)$, where $\text{sim}(\cdot,\cdot)$ is the cosine similarity.

Diffusion Alignment as Sampling (DAS) Aligning pre-trained diffusion models with specific rewards while preserving their generative quality and diversity is a key challenge [9, 7, 12, 49, 42, 6, 14, 72]. Fine-tuning can lead to reward over-optimization, while simpler guidance methods may underperform. [21] proposed DAS, a *training-free* algorithm to address this. DAS aims to effectively sample from a reward-aligned target distribution without explicit model retraining, thereby mitigating over-optimization. The core of DAS is to sample from the target distribution $p_{tar}(x)$, which balances a reward function $r(x)$ with fidelity to the pre-trained model $p_{pre}(x)$:

$$P_{tar}(x) = \frac{1}{\mathcal{Z}} p_{pre}(x) \exp\left(\frac{r(x)}{\alpha}\right), \tag{2}$$

where $\mathcal{Z}$ is a normalization constant and α is a trade-off parameter. To achieve this, DAS employs Sequential Monte Carlo (SMC) [36]. It iteratively guides a set of particles (noisy samples) through the reverse diffusion process. A key feature of DAS is the use of tempered intermediate target distributions $\pi_t(x_t)$ at each diffusion timestep t:

$$\pi_t(x_t) \propto p_t(x_t) \exp\left(\frac{\lambda_t}{\alpha} \hat{r}(x_t)\right), \tag{3}$$

where $p_t(x_t)$ is the marginal distribution from the pre-trained model, $\hat{r}(x_t)$ is the predicted reward from noisy sample x_t, and λ_t is an annealing schedule ($\lambda_T = 0$ to $\lambda_0 = 1$). Diverging from the DAS which predominantly utilized rewards such as HPSv2 [63] and TCE [18] to assess image quality, our approach, in order to specifically address the FSCIL task, involves inputting the generated image $\boldsymbol{x}_{\text{gen}}$ into the classifier. Rewards are subsequently computed based on the classifier's output, and this feedback signal is enhanced by combining multiple distinct rewards.

4 Reward-Aligned Learning via Mutual Boosting Loop for FSCIL

In this section, we introduce our novel framework, Diffusion-Classifier Synergy (DCS), which leverages a reward-aligned learning paradigm through a mutual boosting loop to address the inherent challenges of FSCIL. We first outline the primary issues in FSCIL and present the overarching workflow of our approach. Subsequently, we detail the design of our reward functions at both the feature and logits levels.

4.1 Addressing FSCIL Challenges with a Mutual Boosting Loop

FSCIL faces the *stability-plasticity dilemma* and *unreliable empirical risk minimization*. Generative models, particularly diffusion models, address this by synthesizing data: generating old class images for knowledge replay mitigates the stability-plasticity issue, while creating new class images augments data for few-shot learning. However, naively integrating diffusion models for data generation in FSCIL presents significant problems, which we will analyze in detail.

Semantic Misalignment and Diversity Deficiency of Generated Images. Synthesizing training images that balance semantic fidelity with background diversity presents a substantial challenge when generative prompts are restricted to class labels. More specifically, we employ Stable Diffusion to generate five images per *mini*ImageNet class across various guidance scales, subsequently extracting and classifying features using an ImageNet pre-trained ResNet34. Semantic alignment is determined by classification accuracy. In contrast, intra-class feature dispersion, an indicator of semantic richness quantified by the average L2 distance of features to their class centroid, shows an inverse correlation with accuracy as guidance is strengthened (Figure 2 Left). Critically, irrespective of guidance parameters, synthetic data consistently underperforms on both classification accuracy and semantic richness compared to ResNet34 on the original *mini*ImageNet testset (baseline).

Inefficient Feedback Loop for Guiding Image Generation. Mainstream dataset augmentation methods treat the diffusion model as a *blind* teacher, one that is unaware of the student classifier's needs, which hinders adaptive image generation in FSCIL. Specifically, it limits the ability to produce

samples tailored to the classifier's current capabilities, such as generating images near class decision boundaries. Such fine-grained control is crucial for alleviating semantic confusion between new and old classes. The t-SNE visualization in the right panel of Figure 2 indicates confusing regions between the two classes, which the classifier cannot sufficiently learn from the generated images, leading to easy misclassification of samples in this area.

To overcome these limitations, we propose the Mutual Boosting Loop. Its core idea is to design a multiple reward components $\mathcal{R}_i$, computed based on the output of the classifier σ (with parameters θ) when presented with a generated image x. The diffusion model D is then guided to adjust its sampling strategy ϕ to maximize the sum of these rewards:

$$\phi^* = \arg\max_{\phi} \sum_i \mathcal{R}_i \left(\sigma_\theta(D(x;\phi))\right). \quad (4)$$

The optimized ϕ^* guides image generation $D(x;\phi^*)$, which in turn enhances classifier performance:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{cls}(\sigma_\theta; x \cup D(x;\phi^*), y). \quad (5)$$

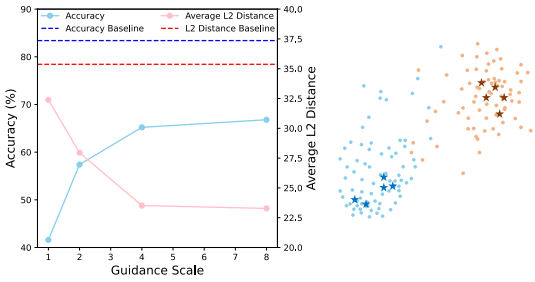

Figure 2: **Left:** Trade-off between semantic fidelity and richness with increasing guidance scale. **Right:** t-SNE visualization of features from real (dot) and generated (pentagram) images illustrates the restricted distribution of the latter within the classifier's decision space.

Consequently, the classifier, now optimized with parameters θ^*, provides more accurate and informative reward signals back to the diffusion model, creating a synergistic co-evolution. The overall process is depicted in Section A. The subsequent sections will elaborate on the construction of this reward mechanism from feature-level (Section 4.2) and logits-level (Section 4.3) perspectives.

4.2 Feature-Level Reward for Semantic Coherence and Diversity

In this section, we detail the process for generating images of a specific class during each session, focusing on feature-level reward components. We assume the classifier has undergone initial learning, with its parameters (class prototypes) initialized using the few available samples for new classes. Addressing the first challenge identified in Section 4.1, a straightforward strategy is to enforce fidelity and diversity using the Fréchet Inception Distance (FID):

$$\text{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}\right), \quad (6)$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. Subscripts r and g indicate real and generated images, respectively. Despite its utility as an evaluation metric, directly using FID as a reward is impractical because restricted old data access prevents its calculation against the comprehensive dataset, and the limited number of new class samples hinders reliable $\boldsymbol{\Sigma}$ estimation, leading to unstable scores.

Drawing inspiration from the principles underlying FID, we propose a more suitable reward mechanism tailored to the constraints of FSCIL. Our goal is to achieve similar objectives of semantic coherence (related to mean matching) and feature diversity (related to covariance matching) using components that are robust with limited data. To approximate the goal of matching the overall feature distribution, particularly ensuring that generated images are semantically anchored to their class identity, we introduce the **Prototype-Anchored Maximum Mean Discrepancy Reward** $\mathcal{R}_{\text{PAMMD}}$. MMD is a statistical test used to determine if two sets of samples are drawn from the same distribution. Instead of directly comparing means which can be problematic with incomplete data for $\boldsymbol{\mu}_r$ and unstable $\boldsymbol{\mu}_g$ from few generated samples, MMD allows for a more holistic comparison.

When a candidate image $\boldsymbol{x}_{\text{gen}}$ is considered for addition to an existing set of $N-1$ generated images $\mathcal{I}_{\text{gen}}^{(c,N-1)}$ of class y_c , $\mathcal{R}_{\text{PAMMD}}$ evaluates the quality of the augmented set $\mathcal{I}_{\text{gen}}^{(c,N)} = \mathcal{I}_{\text{gen}}^{(c,N-1)} \cup \{\boldsymbol{x}_{\text{gen}}\}$ by taking the negative of the MMD:

$$\mathcal{R}_{\text{PAMMD}}(\boldsymbol{x}_{\text{gen}}, \mathcal{I}_{\text{gen}}^{(c,N)}) = -\underbrace{\alpha \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(\boldsymbol{z}_i, \boldsymbol{z}_j)}_{\text{Diversity}} + \underbrace{\beta \frac{1}{N}\sum_{i=1}^{N} k(\boldsymbol{z}_i, \boldsymbol{\mu}_c)}_{\text{Consistency}} - \underbrace{k(\boldsymbol{\mu}_c, \boldsymbol{\mu}_c)}_{\text{Constant}}, \quad (7)$$

https://doi.org/10.52202/085713-2153

where $\boldsymbol{z}_i = f(x_i)$ is the feature of the i-th generated image in $\mathcal{I}_{\text{gen}}^{(c,N)}$. $\boldsymbol{\mu}_c \in \mathbb{R}^D$ is the feature prototype for class y_c, which is in fact the class weight of the classifier. $k(\cdot,\cdot)$ is a positive definite kernel function, and α, β are non-negative hyperparameters.

The first term *Diversity* effectively measures the internal similarity of the $\mathcal{I}_{\text{gen}}^{(c,N)}$. A lower value for this component (contributing to a higher $\mathcal{R}_{\text{PAMMD}}$) is desirable, as it indicates that the generated images are distinct from each other, thereby maximizing diversity. The second term *Consistency* assesses the collective similarity of the generated images to the class prototype and encourages the $\boldsymbol{x}_{\text{gen}}$ to maintain strong semantic consistency with the target class. The third term is a constant that depends only on the class prototype, which can be ignored in the actual computation process. As diffusion models generate images sequentially, the gradient of the $\mathcal{R}_{\text{PAMMD}}$ is backpropagated solely to the currently generated image. To mitigate resource wastage from redundant computations, Equation (7) can be updated incrementally. The detailed procedure is provided in Section D.

A key advantage of $\mathcal{R}_{\text{PAMMD}}$ is its universality. It is applicable whether the model is generating images for previously learned classes or newly introduced classes. As the classifier continuously updates the knowledge acquired for the current class, the prototype $\boldsymbol{\mu}_c$ is also updated correspondingly, enabling it to provide more accurate feedback to the diffusion model in the subsequent session.

Despite the significant utility of $\mathcal{R}_{\text{PAMMD}}$, its reliance solely on prototypes when a subset of new-class images is accessible results in the underutilization of valuable information regarding the feature spread. Drawing inspiration from the second term of the FID, which utilizes covariance matrices to reflect the spatial distribution and spread of features, we propose the **Dimension-Wise Variance Matching Reward** $\mathcal{R}_{\text{VM}}$. This component is developed for robustly matching feature variances of novel classes where limited reference data is available and full covariance estimation is unstable.

We first estimate the per-dimension variances from the features extracted from real images $\mathcal{I}_{\text{real}}^{(c)}$. Then, for each new $\boldsymbol{x}_{\text{gen}}$, we evaluate how it affects the per-dimension variances of the $\mathcal{I}_{\text{gen}}^{(c,N)}$, and reward candidates that bring these variances closer to the target reference variances:

$$\mathcal{R}_{\text{VM}}(\boldsymbol{x}_{\text{gen}}, \mathcal{I}_{\text{gen}}^{(c,N)}) = -\sum_{d=1}^{D} \left(v_{\text{gen}}^d - v_{\text{real}}^d\right)^2, \tag{8}$$

where $v_{\text{gen}}^d = \text{Var}(\{f(\boldsymbol{x}_j)^d \mid \boldsymbol{x}_j \in \mathcal{I}_{\text{gen}}\})$ is the sample variance of the d-th dimension of features in $\mathcal{I}_{\text{gen}}$, and v_{real}^d is defined analogously.

The rationale for dimension-wise matching, rather than full covariance matching, stems from the instability of covariance matrix estimation in few-shot scenarios. Estimating a $D \times D$ covariance matrix from very few samples (*e.g.*, $N \ll D$) can lead to highly inaccurate or singular matrices. Matching variances at the dimensional level alleviates this issue by focusing on more robust univariate statistics, though it forgoes capturing inter-dimensional correlations. Given that this matching process still entails certain sample size requirements, we recommend deferring the incorporation of this term into the overall reward until a sufficient number of images (*e.g.*, > 5) are generated.

4.3 Logits-Level Rewards for Robust Discrimination Learning

While feature-level rewards ensure semantic integrity and diversity, they do not directly leverage the classifier's decision-making process. To bridge this gap and enable finer-grained control over image generation, we introduce logits-level rewards that incorporate feedback from the current classifier.

A fundamental logits-level reward aims to encourage the generation of images that the classifier can correctly assign to the target class y_c. This can be initially conceptualized with a standard cross-entropy reward:

$$\mathcal{R}_{\text{CE}}(\boldsymbol{x}_{\text{gen}}, y_c) = \log(\hat{p}(y_c|\boldsymbol{x}_{\text{gen}})) = \log\left(\frac{\exp(\sigma_c(\boldsymbol{x}_{\text{gen}}))}{\sum_k \exp(\sigma_k(\boldsymbol{x}_{\text{gen}}))}\right), \tag{9}$$

where $\sigma_c(\boldsymbol{x}_{\text{gen}})$ is the logit for the target class y_c, $\hat{p}(y_c|\boldsymbol{x}_{gen})$ is the probability assigned by the current classifier to the target class y_c for the generated image $\boldsymbol{x}_{\text{gen}}$. However, relying solely on maximizing this basic classification accuracy can lead to the generation of overly simplistic or peaky samples. To address this limitation and encourage the generation of more nuanced and informative samples, we

refine this concept into a **Recalibrated Confidence Reward** with temperature-scaled probability:

$$\mathcal{R}_{\text{RC}}(\boldsymbol{x}_{\text{gen}}, y_c) = \log(\hat{p}(y_c|\boldsymbol{x}_{\text{gen}}; T)), \tag{10}$$

$$\hat{p}(y_c|\boldsymbol{x}_{\text{gen}}; T) = \frac{\exp(\sigma_c(\boldsymbol{x}_{\text{gen}})/T)}{\sum_k \exp(\sigma_k(\boldsymbol{x}_{\text{gen}})/T)}, \tag{11}$$

$$T(\boldsymbol{x}_{\text{gen}}) = T_{base} + T_{scale} \cdot \left(\frac{\hat{p}_c(y_c|\boldsymbol{x}_{\text{gen}}) - 1/N_c}{1 - 1/N_c} \right), \tag{12}$$

where N_c is the total class count. $T_{base} > 1$ provides baseline smoothing, and $T_{scale} > 0$ sets the maximum additional temperature. This adaptive temperature adjusts based on the classifier's raw confidence in the target class y_c for $\boldsymbol{x}_{\text{gen}}$. A high $\hat{p}_c(y_c|\boldsymbol{x}_{\text{gen}})$ increases temperature T, flattening the reward's probability distribution to discourage overly simplistic samples. Conversely, a low $\hat{p}_c(y_c|\boldsymbol{x}_{\text{gen}})$ keeps T near T_{base}, minimizing extra smoothing to prevent excessive diffusion. This efficient mechanism dynamically controls classifier feedback smoothness within the logits-level reward, thereby fostering the generation of more challenging and diverse samples.

While the R_{RC} focuses on the confidence of the target class, it does not explicitly address the relationships between different classes. In FSCIL settings, a significant challenge arises when feature embeddings of new classes closely overlap with those of previously learned classes. Simply ensuring a high (even if smoothed) probability for the target class y_c might not be sufficient to explicitly create a clear distinction from these confusing old classes. To this end, we introduce the **Cross-Session Confusion-Aware Reward** $\mathcal{R}_{CSCA}$, which makes it possible to generate *harder* samples for robust training and refine the classifier's understanding of nuanced class differences. To detail how this reward is formulated, we first outline how the classifier assesses these inter-class similarities and potential confusions.

The classifier's decision process involves computing the cosine similarity $\hat{s}(y_c|\boldsymbol{x}_{\text{gen}})$ between the features $f(\boldsymbol{x}_{\text{gen}})$ and the class prototype $\boldsymbol{\mu}_c$ for each class y_c, which is then used to derive the logits:

$$\hat{s}(y_c|\boldsymbol{x}_{\text{gen}}) = \frac{f(\boldsymbol{x}_{\text{gen}}) \cdot \boldsymbol{\mu}_c}{\|f(\boldsymbol{x}_{\text{gen}})\| \|\boldsymbol{\mu}_c\|}. \tag{13}$$

Based on this similarity, the cosine distance is $d_{\cos}(\boldsymbol{x}_{\text{gen}}, \boldsymbol{\mu}_c) = 1 - \hat{s}(y_c|\boldsymbol{x}_{\text{gen}})$, which quantifies how dissimilar the generated sample's features are from the class center $\boldsymbol{\mu}_c$. To modulate the influence of different classes based on their similarity to the generated sample, we introduce dynamic weights $w_{y_t}(\boldsymbol{x}_{\text{gen}})$ such that the weight for a class y_t increases as the sample becomes more similar to the $\boldsymbol{\mu}_t$:

$$w_{y_t}(\boldsymbol{x}_{\text{gen}}) = \frac{1}{1 + \gamma \cdot d_{\cos}(\boldsymbol{x}_{\text{gen}}, \boldsymbol{\mu}_t)}, \tag{14}$$

where the scaling factor $\gamma > 0$ controls the sensitivity of the weight to the cosine distance. A smaller $d_{\cos}(\boldsymbol{x}_{\text{gen}}, y_t)$ result in more weight being assigned to the target class y_t for the generated sample. Leveraging these weights, the $\mathcal{R}_{CSCA}$ is designed to encourage the generator to create samples for a target class y_c that are highly similar to a confusable class y_t within a set $\mathcal{C}$. This is achieved by defining the reward as the log-probability of classifying $\boldsymbol{x}_{\text{gen}}$ as y_t instead of y_c, using dynamically weighted similarity scores as logits:

$$\mathcal{R}_{\text{CSCA}}(\boldsymbol{x}_{\text{gen}}, y_c) = \sum_{y \in \mathcal{C}} w_y(\boldsymbol{x}_{\text{gen}}) \log\left(\hat{p}(y|\boldsymbol{x}_{\text{gen}}; T_s)\right), \tag{15}$$

where $\mathcal{C}$ can be the set of the top-K most similar prototypes $\boldsymbol{\mu}_y$ to $\boldsymbol{x}_{\text{gen}}$ in order to reduce computation.

5 Experiments

5.1 Experimental Setup

Datasets Following the benchmark settings of previous methods, we conducted experiments on three datasets, *i.e.*, *mini*ImageNet [55, 45], CUB-200 [57], and CIFAR-100 [22]. The division of the datasets aligns with existing methods. Specifically, the CIFAR-100 and *mini*ImageNet datasets are partitioned into a base session containing 60 classes and incremental sessions containing 40 classes, with each session being an 8-way 5-shot few-shot classification task. The CUB-200 dataset is divided into the base session containing 100 classes and incremental sessions containing 40 classes, with each session being a 10-way 5-shot task.

https://doi.org/10.52202/085713-2153

Table 1: Comparison results on *mini*ImageNet dataset. * denotes results from [27].

Methods	Accuracy in each session (%) ↑									Avg ↑
	0	1	2	3	4	5	6	7	8	
TOPIC [52]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.64
CEC [73]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	57.75
FACT [77]	72.56	69.63	66.38	62.77	60.60	57.33	54.34	52.16	50.49	60.70
TEEN [60]	73.53	70.55	66.37	63.23	60.53	57.95	55.24	53.44	52.08	61.44
SAVC [50]	81.12	76.14	72.43	68.92	66.48	62.95	59.92	58.39	57.11	67.05
DyCR [38]	73.18	70.16	66.87	63.43	61.18	58.79	55.00	52.87	51.08	61.40
ALFSCIL [26]	81.27	75.97	70.97	66.53	63.46	59.95	56.93	54.81	53.31	64.80
OrCo* [3]	**83.22**	74.60	71.89	67.65	65.53	62.73	60.33	58.51	57.62	66.90
ADBS* [27]	81.40	75.03	71.03	68.00	65.56	61.87	59.04	56.87	55.38	66.02
DCS(Ours)	82.43	**77.54**	**73.00**	**69.21**	**67.05**	**64.44**	**61.20**	**60.43**	**57.99**	**68.14**

Table 2: Comparison results on CUB-200 dataset. * denotes results from [27].

Methods	Accuracy in each session (%) ↑											Avg ↑
	0	1	2	3	4	5	6	7	8	9	10	
TOPIC [52]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28	43.92
CEC [73]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	61.33
FACT [77]	75.90	73.23	70.84	66.13	65.56	62.15	61.74	59.83	58.41	57.89	56.94	64.42
TEEN [60]	77.26	76.13	72.81	68.16	67.77	64.40	63.25	62.29	61.19	60.32	59.31	66.63
SAVC [50]	81.85	**77.92**	**74.95**	70.21	**69.96**	67.02	**66.16**	65.30	63.84	63.15	62.50	69.35
DyCR [38]	77.50	74.73	71.69	67.01	66.59	63.43	62.66	61.69	60.57	59.69	58.46	65.82
ALFSCIL [26]	79.79	76.53	73.12	69.02	67.62	64.76	63.45	62.32	60.83	60.21	59.30	67.00
OrCo* [3]	74.58	65.99	64.72	63.06	61.79	59.55	59.21	58.46	56.97	57.99	57.32	61.79
ADBS* [27]	79.99	75.89	72.53	68.33	67.92	64.75	64.10	62.93	61.31	60.88	59.65	67.12
DCS(Ours)	**83.19**	77.32	73.92	**70.52**	69.79	**68.00**	65.22	**66.59**	**64.19**	**64.86**	**63.40**	**69.73**

Evaluation metrics Session accuracy quantifies model performance within a specific learning session. To evaluate sustained performance and generalization across both previously learned and newly introduced classes, average accuracy is calculated as the mean of all session accuracies from the initial to the current session.

Implementation details For the Diffusion Alignment as Sampling (DAS) algorithm, we utilized the source code made available by the authors, into which our proposed reward mechanisms were integrated. The latest Stable Diffusion 3.5 Medium model [11] served as the foundational diffusion model. To align with the configuration of Stability AI's open-source weights, thereby ensuring the fidelity of the generated images, we adapted the DAS source code for compatibility with Flow Matching Scheduler. Images were generated at a resolution of 512×512 pixels and subsequently resized to match the native resolution of the real dataset before being input to the encoder. During the base session, an additional 30 images are generated per class. For new sessions, we generate 30 and 10 images for each newly introduced and previously learned class, respectively. For more details, please refer to Section B.

5.2 Comparison with the State of the Art

In this section, we compare our proposed DCS with mainstream methods on FSCIL benchmarks. Table 1 and Table 2 present the accuracy in each session and the average accuracy of these methods on the *mini*ImageNet and CUB-200, respectively.

Compared to previous state-of-the-art FSCIL methods, our proposed DCS achieves the highest accuracy in each session. Note that the compared methods employ network optimization techniques tailored to the characteristics of FSCIL tasks, including additional self-supervised learning or distribution calibration. In contrast, our DCS achieves performance improvement solely through the generalized knowledge derived from the diffusion model, without any modifications to the baseline classification network. For further results, please refer to Section C.

Table 3: Ablation studies on CIFAR-100 benchmark. Δ_{last}: Relative improvements of the last sessions compared to the baseline.

$\mathcal{R}_{\text{PAMMD}}$	$\mathcal{R}_{\text{VM}}$	$\mathcal{R}_{\text{RC}}$	$\mathcal{R}_{\text{CSCA}}$	Accuracy in each session (%) ↑								Δ_{last}
				1	2	3	4	5	6	7	8	
Baseline				71.97	67.55	62.83	59.73	56.08	52.99	51.55	49.57	-
✓				73.90	69.05	64.70	60.99	57.91	55.02	53.82	50.81	+1.24
✓	✓			74.05	69.29	65.06	61.50	58.59	55.63	54.61	51.43	+1.86
✓	✓	✓		75.08	70.77	66.44	62.94	60.61	57.47	57.01	53.07	+3.50
✓	✓	✓	✓	**75.96**	**72.06**	**67.35**	**64.38**	**62.12**	**60.05**	**58.99**	**55.21**	**+5.64**

5.3 Further Analysis

Ablation Study An ablation study was conducted to investigate the contribution of these components, and Table 3 summarizes their performance on the CIFAR-100 dataset. To preclude interference from performance disparities in the base session with the experimental results, we employ fixed weights derived from the base session and initiate experiments starting from the first new session. The contribution of $\mathcal{R}_{\text{PAMMD}}$ in aligning semantics yields an accuracy improvement of 1.24% compared to the baseline, an advantage further extended to 1.86% by $\mathcal{R}_{\text{VM}}$. At the logits level, $\mathcal{R}_{\text{RC}}$ demonstrates a more pronounced contribution to accuracy enhancement compared to the feature-level rewards, boosting performance from 1.86% to 3.50%. This underscores the critical role of feedback from the classifier's decision space in improving the training effectiveness of the generated images. $\mathcal{R}_{\text{CSCA}}$ further highlights the efficacy of generating customized hard samples based on the classifier's mastery of the training data, leading to a substantial accuracy increase to 5.64%.

The Diffusion Model's Intrinsic Baseline Using pre-trained diffusion models for downstream tasks requires careful consideration of their inherent knowledge's impact on results. To investigate this, we conducted experiments by generating varying quantities of training images, employing only a foundational CLIP-score reward while maintaining consistency across other parameters. Figure 3 illustrates the average classification accuracy in new sessions under different guidance scales. While empirical evidence suggests that a larger volume of generated samples correlates with improved training efficacy, the substantial resource expenditure associated with such large-scale generation is incongruent with the practical constraints of FSCIL scenarios. Particularly in low-data regimes (*i.e.*, fewer than 50 generated images per class), the performance of baseline methods fails to reach the level achieved by our proposed approach (indicated by the dashed line). DCS offers a approach that enables achieving performance comparable to that obtained with a larger volume of images, even when using a minimal number of generated images under resource-constrained conditions.

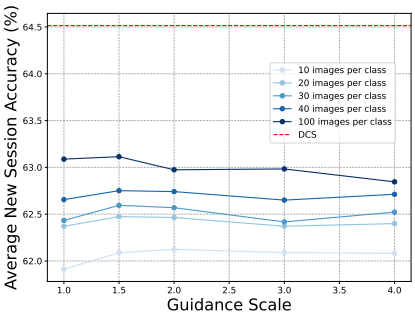

Figure 3: **Training performance comparison.** DCS versus the vanilla diffusion model generating images at varying quantities and guidance scales, with the latter underperforming at comparable generation scales.

6 Conclusion

This paper introduced DCS, a novel framework advancing FSCIL by establishing a mutual boosting loop between a diffusion model and a classifier through reward-aligned learning. DCS employs a feature-level as well as logits-level reward system to guide the generation of strategically beneficial images, ensuring they are tailored to the classifier's evolving needs. Empirical validation on benchmark datasets demonstrated DCS's state-of-the-art performance in mitigating catastrophic forgetting and effectively learning new classes from minimal data. The study confirmed that aligning generative processes with classifier goals is crucial for continual learning, emphasizing DCS's role in using diffusion models to address major FSCIL challenges and enhance adaptability.

https://doi.org/10.52202/085713-2153

Acknowledgments

This work is partially supported by grants from the National Natural Science Foundation of China (No. 62132002, No. 62202010, and No. 62402015), the Beijing Nova Program (No. 20250484786), the China Postdoctoral Science Foundation under grant (No. 2024M750100), the Postdoctoral Fellowship Program of CPSF under grant (No. GZB20230024), and the Fundamental Research Funds for the Central Universities.

References

[1] Diffusers — huggingface.co. https://huggingface.co/docs/diffusers/en/index.

[2] Aishwarya Agarwal, Biplab Banerjee, Fabio Cuzzolin, and Subhasis Chaudhuri. Semantics-driven generative replay for few-shot class incremental learning. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 5246–5254. ACM, 2022.

[3] Noor Ahmed, Anna Kukleva, and Bernt Schiele. Orco: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 28762–28771. IEEE, 2024.

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Trans. Mach. Learn. Res.*, 2023, 2023.

[5] Jihwan Bang, Heesu Kim, Youngjoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8218–8227. Computer Vision Foundation / IEEE, 2021.

[6] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[7] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[8] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14146–14155. IEEE, 2022.

[9] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[10] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[12] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[13] Yunxiang Fu, Chaoqi Chen, Yu Qiao, and Yizhou Yu. Dreamda: Generative data augmentation with diffusion models. *CoRR*, abs/2403.12803, 2024.

[14] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[15] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9047–9057. IEEE, 2022.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in*

Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[17] Zitong Huang, Ze Chen, Zhixing Chen, Erjin Zhou, Xinxing Xu, Rick Siow Mong Goh, Yong Liu, Chun-Mei Feng, and Wangmeng Zuo. Learning prompt with distribution-based feature replay for few-shot class-incremental learning. *CoRR*, abs/2401.01598, 2024.

[18] Francisco Ibarrola and Kazjon Grace. Measuring diversity in co-creative image generation. In Kazjon Grace, Maria Teresa Llano, Pedro Martins, and Maria M. Hedblom, editors, *Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024*, pages 254–261. Association for Computational Creativity (ACC), 2024.

[19] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, Karthik Nandakumar, and Naveed Akhtar. Genmix: Effective data augmentation with generative diffusion model image editing. *CoRR*, abs/2412.02366, 2024.

[20] Haeyong Kang, Jaehong Yoon, Sultan Rizky Hikmawan Madjid, Sung Ju Hwang, and Chang D. Yoo. On the soft-subnetwork for few-shot class incremental learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[21] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[23] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8230–8239. IEEE, 2021.

[24] Hanbyul Lee and Juneho Yi. Few-shot class-incremental model attribution using learnable representation from clip-vit features. *ArXiv*, abs/2503.08148, 2025.

[25] Janghyeon Lee, Hyeong Gwon Hong, Donggyu Joo, and Junmo Kim. Continual learning with extended kronecker-factored approximate curvature. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8998–9007. Computer Vision Foundation / IEEE, 2020.

[26] Jiashuo Li, Songlin Dong, Yihong Gong, Yuhang He, and Xing Wei. Analogical learning-based few-shot class-incremental learning. *IEEE Trans. Circuits Syst. Video Technol.*, 34(7):5493–5504, 2024.

[27] Linhao Li, Yongzhang Tan, Siyuan Yang, Hao Cheng, Yongfeng Dong, and Liang Yang. Adaptive decision boundary for few-shot class-incremental learning. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 18359–18367. AAAI Press, 2025.

[28] Xiaojie Li, Yibo Yang, Xiangtai Li, Jianlong Wu, Yue Yu, Bernard Ghanem, and Min Zhang. Genview: Enhancing view quality with pretrained generative model for self-supervised learning. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVIII*, volume 15126 of *Lecture Notes in Computer Science*, pages 306–325. Springer, 2024.

[29] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3925–3934. PMLR, 2019.

[30] Grace Luo, Trevor Darrell, Oliver Wang, Dan B. Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8217–8227. IEEE, 2024.

[31] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5513–5533, 2023.

[32] M. Anwar Ma'sum, Mahardhika Pratama, and Igor Skrjanc. Latest advancements towards catastrophic forgetting under data scarcity: A comprehensive survey on few-shot class incremental learning. *CoRR*, abs/2502.08181, 2025.

[33] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2624–2637, 2013.

[34] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4, 2013.

[35] Eyal Michaeli and Ohad Fried. Advancing fine-grained classification by structure and subject preserving augmentation. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[36] Pierre Del Moral and A. Doucet. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 2002.

[37] Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Closer: Towards better representation learning for few-shot class-incremental learning. In *European Conference on Computer Vision*, 2024.

 https://doi.org/10.52202/085713-2153

[38] Zicheng Pan, Xiaohan Yu, Miaohua Zhang, Weichuan Zhang, and Yongsheng Gao. Dycr: A dynamic clustering and recovering network for few-shot class-incremental learning. *IEEE Trans. Neural Networks Learn. Syst.*, 36(4):7116–7129, 2025.

[39] Kirill Paramonov, Mete Ozay, Eunju Yang, Ji Joong Moon, and Umberto Michieli. Controllable forgetting mechanism for few-shot class-incremental learning. *ArXiv*, abs/2501.15998, 2025.

[40] Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 23881–23890. IEEE, 2024.

[41] Can Peng, Kun Zhao, Tianren Wang, Meng Li, and Brian C. Lovell. Few-shot class-incremental learning from an open-set perspective. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXV*, volume 13685 of *Lecture Notes in Computer Science*, pages 382–397. Springer, 2022.

[42] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *ArXiv*, abs/2310.03739, 2023.

[43] Hang Ran, Weijun Li, Lusi Li, Songsong Tian, Xin Ning, and Prayag Tiwari. Learning optimal inter-class margin adaptively for few-shot class-incremental learning via neural collapse-based meta-learning. *Inf. Process. Manag.*, 61(2):103664, 2024.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[46] Minhyuk Seo, Diganta Misra, Seongwon Cho, Minjae Lee, and Jonghyun Choi. Just say the name: Online continual learning with category names only via data generation. *CoRR*, abs/2403.10853, 2024.

[47] Abhilash Reddy Shankarampeta and Koichiro Yamauchi. Few-shot class incremental learning with generative feature replay. In Maria De Marsico, Gabriella Sanniti di Baja, and Ana Fred, editors, *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2021, Online Streaming, February 4-6, 2021*, pages 259–267. SCITEPRESS, 2021.

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[49] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 32483–32498. PMLR, 2023.

[50] Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24183–24192. IEEE, 2023.

[51] Yu-Ming Tang, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Rethinking few-shot class-incremental learning: Learning from yourself. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXI*, volume 15119 of *Lecture Notes in Computer Science*, pages 108–128. Springer, 2024.

[52] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12180–12189. Computer Vision Foundation / IEEE, 2020.

[53] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2024.

[54] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[55] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638, 2016.

[56] Anh-Khoa Nguyen Vu, Quoc-Truong Truong, Vinh-Tiep Nguyen, Thanh Duc Ngo, Thanh-Toan Do, and Tam V. Nguyen. Multi-perspective data augmentation for few-shot object detection. *CoRR*, abs/2502.18195, 2025.

[57] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[58] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: feature boosting and compression for class-incremental learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXV*, volume 13685 of *Lecture Notes in Computer Science*, pages 398–414. Springer, 2022.

[59] Liyuan Wang, Kuo Yang, Chongxuan Li, Lanqing Hong, Zhenguo Li, and Jun Zhu. Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5383–5392. Computer Vision Foundation / IEEE, 2021.

[60] Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[61] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), June 2020.

[62] Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 17223–17233. IEEE, 2024.

[63] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2096–2105. IEEE, 2023.

[64] Ju Xu and Zhanxing Zhu. Reinforced continual learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 907–916, 2018.

[65] Shipeng Yan, Jiangwei Xie, and Xuming He. DER: dynamically expandable representation for class incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3014–3023. Computer Vision Foundation / IEEE, 2021.

[66] Boyu Yang, Mingbao Lin, Binghao Liu, Mengying Fu, Chang Liu, Rongrong Ji, and Qixiang Ye. Learnable expansion-and-compression network for few-shot class-incremental learning. *CoRR*, abs/2104.02281, 2021.

[67] Boyu Yang, Mingbao Lin, Yunxiao Zhang, Binghao Liu, Xiaodan Liang, Rongrong Ji, and Qixiang Ye. Dynamic support network for few-shot class incremental learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):2945–2951, 2023.

[68] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip H. S. Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[69] Yang Yang, Da-Wei Zhou, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Adaptive deep models for incremental learning: Considering capacity scalability and sustainability. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 74–82. ACM, 2019.

[70] Yang Yang, Da-Wei Zhou, De-Chuan Zhan, Hui Xiong, Yuan Jiang, and Jian Yang. Cost-effective incremental deep model: Matching model capacity with the least sampling. *IEEE Trans. Knowl. Data Eng.*, 35(4):3575–3588, 2023.

[71] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[72] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23117–23127. IEEE, 2023.

[73] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12455–12464. Computer Vision Foundation / IEEE, 2021.

[74] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13205–13214. Computer Vision Foundation / IEEE, 2020.

[75] Hanbin Zhao, Hui Wang, Yongjian Fu, Fei Wu, and Xi Li. Memory-efficient class-incremental learning for image classification. *IEEE Trans. Neural Networks Learn. Syst.*, 33(10):5966–5977, 2022.

[76] Yifan Zhao, Jia Li, Zeyin Song, and Yonghong Tian. Language-inspired relation transfer for few-shot class-incremental learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2):1089–1102, 2025.

 https://doi.org/10.52202/085713-2153

[77] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9036–9046. IEEE, 2022.

[78] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9851–9873, 2024.

[79] Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[80] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):12816–12831, 2023.

[81] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Class-incremental learning via dual augmentation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14306–14318, 2021.

[82] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5871–5880. Computer Vision Foundation / IEEE, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly pointed out the problem we studied (FSCIL) in the abstract, and listed our main contributions in detail in the introduction. We also verified our contribution through extensive analysis and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the appendix of this paper (in the supplementary material), we discuss the limitations of the proposed approach, with a primary focus on the usability of diffusion models.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

https://doi.org/10.52202/085713-2153

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not present theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a detailed description of the experimental setup, including the three benchmark datasets utilized, the version of the diffusion model, and the various hyperparameters used throughout the training process. Based on these descriptions, readers can be able to easily replicate the experimental results presented in this paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We commit to publicly releasing the source code immediately upon acceptance of our paper. It is important to emphasize that all datasets used in the paper are public. The division method and testing procedures for the FSCIL task, as discussed, are all open-source, and our experimental process is in full compliance with them. As stated in Question 4 (Experimental Result Reproducibility), we have provided a sufficiently detailed description of the experimental process and parameter settings. We did not employ any additional tricks, which makes our work easily reproducible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide comprehensive details of the experiments. The experiment section in the main paper briefly introduces the dataset and the basic settings, while the supplementary materials provide an in-depth description of all the details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

https://doi.org/10.52202/085713-2153

Justification: In line with many prior works in the FSCIL literature, we report mean accuracies over multiple runs but do not include error bars or formal statistical significance tests. We acknowledge this as a limitation. However, we argue for the significance of our results based on two factors: (1) the substantial and consistent margin of improvement over state-of-the-art methods across all sessions and three distinct benchmarks, which makes random fluctuation an unlikely explanation; and (2) our extensive ablation studies (Table 3, Table 5) that demonstrate the clear, positive contribution of each component of our method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the supplementary material, we provide a detailed description of the experimental environment, including hardware resources and software versions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have conducted a thorough review of the NeurIPS Ethical Guidelines and commit that the research presented in our paper adheres to these guidelines in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential social implications of our work in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not propose a generative model. The principal model employed in this paper is ResNet, which serves as a feature extractor and does not present such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

https://doi.org/10.52202/085713-2153

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All data and models used in this paper are open source and accessible, and we have cited all sources accordingly.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper has not yet released the model weights and code. However, we commit to providing well-organized files and clear, user-friendly documentation when these are made available as open-source resources.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA] .

    Justification: The core method development in this study does not involve any LLM.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

https://doi.org/10.52202/085713-2153

A Overall Pipeline

DCS aims to enhance the generalization ability of the classifier by constructing bidirectional feedback between the diffusion model and the classifier, while only generating a few images. The workflow of DCS is detailed in Algorithm 1. Compared to the original FSCIL learning strategy, DCS only adds the step of generating new images.

In the base session, we first train the classifier using real data. Since the data in this session is sufficient, we focus on generating more challenging samples with higher ambiguity (by adding $\mathcal{R}_{\text{CSCA}}$ based on the feature-level reward) and use them to fine-tune the classifier.

In each incremental session, we first initialize the classifier weights W_c using the limited real data from the new classes (as described in Section 3). Then, we generate additional learning data for both the new and previously learned old classes. For the new classes, to prevent overfitting, DCS aims to help the classifier quickly build an understanding of their concepts. Therefore, we only use $\mathcal{R}_{\text{RC}}$ at the logits level. For the old classes, the focus is on their confusion with the new classes, so $\mathcal{R}_{\text{CSCA}}$ is applied, as in the base session. Finally, we mix the newly generated data for both new and old classes and fine-tune the classifier.

Algorithm 1 FSCIL with DCS

Data: $\mathcal{D}_{base}$, $\mathcal{D}^t_{inc}$ (K-shot, for $t = 1, \ldots, T$)
Models:
  Classifier $\sigma = (f, W)$ with parameters (θ_f, θ_W)
  Diffusion Model D
Loss: $\mathcal{L}_{cls}$
Rewards: $\mathcal{R}_{\text{PAMMD}}, \mathcal{R}_{\text{VM}}, \mathcal{R}_{\text{RC}}, \mathcal{R}_{\text{CSCA}}$
Generated Images Count: $N_{base}, N_{new}, N_{old}$ per class

// Base Session
$\mathcal{C}_{base} \leftarrow$ classes in $\mathcal{D}_{base}$
$(\theta_f, \theta_W) \leftarrow \arg\min_{\theta_f, \theta_W} \mathcal{L}_{cls}(\sigma(\mathcal{D}_{base}), \mathcal{C}_{base})$
$X_{gen\_base} \leftarrow \varnothing$
$\mathcal{R}_{base} \leftarrow \mathcal{R}_{\text{PAMMD}} + \mathcal{R}_{\text{VM}} + \mathcal{R}_{\text{CSCA}}$
for all class $c \in \mathcal{C}_{base}$ **do** ▷ Generate images for the base classes
  $X^c_{gen\_base} \leftarrow \text{Generate}(D, c, N_{base}, \mathcal{R}_{base}, \sigma)$
  $X_{gen\_base} \leftarrow X_{gen\_base} \cup X^c_{gen\_base}$
$(\theta_f, \theta_W) \leftarrow \arg\min_{\theta_f, \theta_W} \mathcal{L}_{cls}(\sigma(\mathcal{D}_{base} \cup X_{gen\_base}), \mathcal{C}_{base})$ ▷ Fine-tune classifier
$\mathcal{C}_{seen} \leftarrow \mathcal{C}_{base}$

// Incremental Session
for all session $t = 1, \ldots, T$ **do**
  $\mathcal{C}^t_{new} \leftarrow$ classes in $\mathcal{D}^t_{inc}$
  $\mathcal{C}^t_{old} \leftarrow \mathcal{C}_{seen}$
  $\theta_W \leftarrow \frac{1}{K} \sum_{i=1}^{K} f(\mathcal{D}^t_{inc,i})$ ▷ Initialize θ_W for $\mathcal{C}^t_{new}$ using $\mathcal{D}^t_{inc}$
  $X_{gen\_inc} \leftarrow \varnothing$
  $\mathcal{R}_{new} \leftarrow \mathcal{R}_{\text{PAMMD}} + \mathcal{R}_{\text{VM}} + \mathcal{R}_{\text{RC}}$
  for all class $c \in \mathcal{C}^t_{new}$ **do** ▷ Generate images for the new classes
    $X^c_{gen\_inc} \leftarrow \text{Generate}(D, c, N_{new}, \mathcal{R}_{new}, \sigma)$
    $X_{gen\_inc} \leftarrow X_{gen\_inc} \cup X^c_{gen\_inc}$
  $\mathcal{R}_{old} \leftarrow \mathcal{R}_{\text{PAMMD}} + \mathcal{R}_{\text{VM}} + \mathcal{R}_{\text{CSCA}}$
  for all class $c \in \mathcal{C}^t_{old}$ **do** ▷ Generate images for the old classes
    $X^c_{gen\_inc} \leftarrow \text{Generate}(D, c, N_{old}, \mathcal{R}_{old}, \sigma)$
    $X_{gen\_inc} \leftarrow X_{gen\_inc} \cup X^c_{gen\_inc}$
  $\mathcal{D}^t_{train\_inc} \leftarrow \mathcal{D}^t_{inc} \cup X_{gen\_inc}$
  $(\theta_f, \theta_W) \leftarrow \arg\min_{\theta_f, \theta_W} \mathcal{L}_{cls}(\sigma(\mathcal{D}^t_{train\_inc}), \mathcal{C}^t_{new} \cup \mathcal{C}^t_{old})$ ▷ Fine-tune classifier
  $\mathcal{C}_{seen} \leftarrow \mathcal{C}_{seen} \cup \mathcal{C}^t_{new}$
  Evaluate $\sigma(\cdot; \theta_f, \theta_W)$ on $\mathcal{C}_{seen}$.

https://doi.org/10.52202/085713-2153

B Implementation Details

Diffusion Model As described in Section 5.1, we use Stable Diffusion 3.5 Medium [11] to generate images. For all datasets, the `guidance_scale` is set to 2.0. The prompt is "a photo of {class name}", and the negative prompt is "Anime, smooth, close-up, virtual, logo, partial magnification, exquisite". The number of steps for each image is set to 10.

Diffusion Alignment as Sampling (DAS) [21] For all datasets, the `kl_coeff` is set to 0.001, `num_particles` is set to 16, and `tempering_gamma` is set to 0.008.

Hyperparameters In $\mathcal{R}_{\text{PAMMD}}$, $\alpha = 1$, $\beta = 2$, and $k = 0$ (since this term is constant). In $\mathcal{R}_{\text{RC}}$, T_{base} is set to 2.0, and T_{scale} is set to 1.0. In $\mathcal{R}_{\text{CSCA}}$, γ is set to 1.0, and only the top-3 old classes y_t most similar to y_c are considered.

Training Details For the optimizer, SGD is used on all datasets with momentum of 0.9 and weight decay of 0.0005. In addition, we use the cosine annealing strategy to dynamically adjust the learning rate during training. For the CIFAR-100 dataset, the initial learning rate is 0.1 with 50 epochs for the base session, and 0.01 with 5 epochs for the incremental session. For the *mini*ImageNet dataset, the initial learning rate is 0.1 with 120 epochs for the base session, and 0.05 with 30 epochs for the incremental session. For the CUB-200 dataset, the initial learning rate is 0.002 with 120 epochs for the base session, and 0.0005 with 10 epochs for the incremental session. Following [15, 68, 27], we employ ResNet-18 as the backbone for CUB200, and ResNet-12 for *mini*ImageNet and CIFAR100. After the base session, we freeze the shallow layers of ResNet and keep only the last layer for training, with its learning rate individually set to 0.001 times the learning rate of the new class listed above.

Experimental Environment All experiments were performed under Ubuntu 20.04.4 LTS operating system with NVIDIA GeForce RTX 4090 GPU. The experimental code is written in Python 3.8.19, and the PyTorch (version 1.13.0+cu117) is used for the deep learning framework. The source code of the diffusion model used in the experiments is from the open source library `diffusers` [1] (version 0.32.2).

C More Results

C.1 Comparison results on the CIFAR-100 Dataset.

In the main paper, we compare different methods in accuracy across all sessions on CUB-200 and *mini*ImageNet. In Table 4, we report the detailed comparison results on CIFAR-100 dataset.

Table 4: Comparison results on CIFAR-100 dataset. * denotes results from [27].

Methods	Accuracy in each session (%) ↑									Avg ↑
	0	1	2	3	4	5	6	7	8	
TOPIC [52]	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37	42.62
CEC [73]	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	59.53
FACT [77]	74.60	72.09	67.56	63.52	61.38	58.36	56.28	54.24	52.10	62.24
TEEN [60]	74.92	72.65	68.74	65.01	62.01	59.29	57.90	54.76	52.64	63.10
SAVC [50]	78.77	73.31	69.31	64.93	61.70	59.25	57.13	55.19	53.12	63.63
DyCR [38]	75.73	73.29	68.71	64.80	62.11	59.25	56.70	54.56	52.24	63.04
ALFSCIL [26]	80.75	**77.88**	**72.94**	**68.79**	**65.33**	**62.15**	60.02	57.68	55.17	**66.75**
OrCo* [3]	79.77	63.29	62.39	60.13	58.76	56.56	55.49	54.19	51.12	60.19
ADBS* [27]	79.93	75.22	71.11	65.99	62.46	58.38	55.96	53.72	51.15	63.77
DCS(Ours)	**81.09**	75.96	72.06	67.35	64.38	62.13	**60.05**	**58.99**	**55.21**	66.36

https://doi.org/10.52202/085713-2153

Table 5: Ablation study on generation quality metrics (FID↓, CLIP Score↑). The *vanilla* baseline uses the diffusion model without any reward guidance. The key insight is the *relative improvement* across settings. The degradation in FID/CLIP from adding R_{CSCA} is by design, as this reward's objective is to generate challenging samples near decision boundaries to improve classifier robustness, not to generate high-fidelity images.

Reward Combination	FID-CF100↓	FID-mini↓	FID-CUB↓	CLIP-CF100↑	CLIP-mini↑	CLIP-CUB↑
w/o reward (vanilla)	142.6	135.8	152.1	68.3	69.5	67.1
R_{PAMMD}	128.7	120.4	139.3	72.3	73.4	70.6
+ R_{VM}	**122.6**	**113.8**	**133.1**	74.8	76.3	72.5
+ R_{VM} + R_{RC}	123.4	114.3	134.3	**78.8**	**80.2**	**76.1**
+ R_{VM} + R_{CSCA}	133.0	124.2	144.4	71.9	72.0	70.2

C.2 Extended Ablation Studies

To complement the ablation study in the main paper, we provide further analyses on the impact of our reward components on generation quality, the necessity of each component, and the generalizability of the reward design across different datasets.

C.2.1 Ablation on Generation Quality Metrics

In addition to downstream task performance, an analysis was conducted on the impact of our reward components on standard generation quality metrics, namely the Fréchet Inception Distance (FID) and CLIP Score. Under the data-scarce conditions of FSCIL, the estimation of such metrics from a limited number of samples (e.g., 5-30) is subject to high variance and may not fully represent the true distribution quality. Accordingly, the following evaluation prioritizes the **relative improvement** across reward configurations over absolute scores, as the latter are not directly comparable to benchmarks in large-scale generation literature. Emphasizing relative gains provides a more precise assessment of the targeted effect of each reward component.

The ablation study was performed on the first class of the first incremental session for CIFAR-100, *mini*ImageNet, and CUB-200. For each experimental setting, 30 images were generated, and the average scores are reported in Table 5. The results indicate that the feature-level rewards, R_{PAMMD} and R_{VM}, significantly improve both FID and CLIP scores by promoting semantic consistency and aligning the feature distribution with real data. The inclusion of R_{RC} leads to the best CLIP scores, as it directly optimizes for classification confidence, pushing generated samples to be more semantically pure. Conversely, adding the confusion-aware reward R_{CSCA} degrades these metrics. This result validates our hypothesis: the goal of R_{CSCA} is to generate strategically challenging samples at the decision boundary, which are by nature less aligned with the clean data distribution. This demonstrates that our reward system can generate not only high-fidelity images but also targeted samples tailored to the classifier's evolving needs.

C.2.2 Ablation on Reward Component Necessity

To demonstrate that all four reward components are necessary for the final performance, we conducted a *leave-one-out* analysis on the CIFAR-100 dataset. As shown in Table 6, removing any single component from the full DCS framework results in a performance decrease, confirming that all components contribute synergistically.

Table 6: Leave-one-out ablation study on CIFAR-100. Removing any component degrades the final average accuracy, demonstrating their synergistic contribution.

Method	Avg. Acc. (%)
DCS (Full Model)	**66.36**
DCS w/o R_{PAMMD}	65.12
DCS w/o R_{VM}	65.54
DCS w/o R_{RC}	64.07
DCS w/o R_{CSCA}	64.68

Table 7: Sequential ablation studies on *mini*ImageNet and CUB-200, showing consistent performance improvement as reward components are added.

***mini*ImageNet**		**CUB-200**	
Method	**Avg. Acc. (%)**	**Method**	**Avg. Acc. (%)**
Baseline (w/o reward)	64.03	Baseline (w/o reward)	65.21
+ R_{PAMMD}	65.28	+ R_{PAMMD}	66.35
+ R_{PAMMD} + R_{VM}	65.57	+ R_{PAMMD} + R_{VM}	66.89
+ R_{PAMMD} + R_{VM} + R_{RC}	66.85	+ R_{PAMMD} + R_{VM} + R_{RC}	67.41
DCS (Full Model)	**68.14**	**DCS (Full Model)**	**69.73**

C.2.3 Generalizability of the Reward Design

To confirm that the effectiveness of our reward design is not confined to a single dataset, we performed the same sequential *add-one* ablation study on the *mini*ImageNet and CUB-200 datasets. The results, presented in Table 7, show a consistent and positive trend in performance gains as each reward component is added, demonstrating the general applicability of our approach across diverse benchmarks.

C.3 Qualitative Analysis of Generated Images

Figure 4 displays a qualitative comparison, where the first column shows original images from the *mini*ImageNet dataset and the remaining columns show examples generated by our method. Guided by text prompts and our multi-faceted reward system, the generated images are not only semantically correct but are also rich in scenic diversity.

D Incremental Computation of Reward Functions

This section details the incremental update rules for the Prototype-Anchored Maximum Mean Discrepancy Reward ($\mathcal{R}_{\text{PAMMD}}$) and the Dimension-Wise Variance Matching Reward ($\mathcal{R}_{\text{VM}}$). These rules allow for efficient computation as the diffusion model generates images sequentially. We denote the feature vector of the k-th generated image as $\boldsymbol{z}_k$.

D.1 Incremental Computation of $\mathcal{R}_{\text{PAMMD}}$

Recall the $\mathcal{R}_{\text{PAMMD}}$ formula for a set of N generated features $\mathcal{I}_{\text{gen}}^{(N)}$ and a class prototype $\boldsymbol{\mu}_c$:

$$\mathcal{R}_{\text{PAMMD}}(\boldsymbol{x}_{\text{gen}}, \mathcal{I}_{\text{gen}}^{(c,N)}) = -\frac{\alpha}{N^2} \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N} k(\boldsymbol{z}_i, \boldsymbol{z}_j)}_{\text{Diversity}} + \frac{\beta}{N} \underbrace{\sum_{i=1}^{N} k(\boldsymbol{z}_i, \boldsymbol{\mu}_c)}_{\text{Consistency}}, \tag{16}$$

We need to incrementally update term *Diversity* and term *Consistency* when a new (N)-th feature $\boldsymbol{z}_N$ is generated and added to a set of $N-1$ existing features.

Let

$$S_D^{(N-1)} = \text{Diversity}_{N-1} = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} k(\boldsymbol{z}_i, \boldsymbol{z}_j), \tag{17}$$

$$S_C^{(N-1)} = \text{Consistency}_{N-1} = \sum_{i=1}^{N-1} k(\boldsymbol{z}_i, \boldsymbol{\mu}_c). \tag{18}$$

(1) Initialization ($N = 0$)

Before any image is generated for the class:

$$S_D^{(0)} = 0, \tag{19}$$

https://doi.org/10.52202/085713-2153

$$S_C^{(0)} = 0. \tag{20}$$

(2) First Sample ($N = 1$)

When the first image feature $\boldsymbol{z}_1$ is generated:

$$S_D^{(1)} = k(\boldsymbol{z}_1, \boldsymbol{z}_1), \tag{21}$$

$$S_C^{(1)} = k(\boldsymbol{z}_1, \boldsymbol{\mu}_c), \tag{22}$$

(3) Subsequent Samples ($N > 1$)

Assume we have $S_D^{(N-1)}$ and $S_C^{(N-1)}$ for $N-1$ samples, and a new feature $\boldsymbol{z}_N$ is generated. The new sum $S_D^{(N)}$ includes the old sum $S_D^{(N-1)}$, the self-similarity of the new sample $k(\boldsymbol{z}_N, \boldsymbol{z}_N)$, and twice the sum of similarities between the new sample and all $N-1$ previous samples (due to symmetry $k(\boldsymbol{z}_N, \boldsymbol{z}_i) + k(\boldsymbol{z}_i, \boldsymbol{z}_N)$ where k is symmetric:

$$S_D^{(N)} = S_D^{(N-1)} + k(\boldsymbol{z}_N, \boldsymbol{z}_N) + 2\sum_{i=1}^{N-1} k(\boldsymbol{z}_N, \boldsymbol{z}_i). \tag{23}$$

The new sum $S_C^{(N)}$ is the old sum $S_C^{(N-1)}$ plus the similarity of the new sample to the prototype:

$$S_C^{(N)} = S_C^{(N-1)} + k(\boldsymbol{z}_N, \boldsymbol{\mu}_c). \tag{24}$$

With $S_D^{(N)}$ and $S_C^{(N)}$, the $\mathcal{R}_{\text{PAMMD}}(\mathcal{I}_{\text{gen}}^{(N)}, \boldsymbol{\mu}_c)$ can be calculated using the Equation (16). To compute $\sum_{i=1}^{N-1} k(\boldsymbol{z}_N, \boldsymbol{z}_i)$, we need access to the feature vectors of the $N-1$ previously accepted images.

D.2 Incremental Computation of $\mathcal{R}_{\text{VM}}$

Recall the $\mathcal{R}_{\text{VM}}$ for a candidate $\boldsymbol{z}_N$ when added to a set of $N-1$ generated features, forming an augmented set $\mathcal{I}_{\text{gen}}^{(N)}$:

$$\mathcal{R}_{\text{VM}}(\boldsymbol{x}_{\text{gen}}, \mathcal{I}_{\text{gen}}^{(c,N)}) = -\sum_{d=1}^{D} \left(v_{\text{gen},[d]} - v_{\text{real},[d]}\right)^2, \tag{25}$$

where $v_{\text{gen},[d]} = \text{Var}(\{f(\boldsymbol{x}_j)^d \mid \boldsymbol{x}_j \in \mathcal{I}_{\text{gen}}\})$ is the sample variance of the d-th dimension of features in $\mathcal{I}_{\text{gen}}$, and $v_{\text{real},[d]}$ is defined analogously.

The sample variance for a set of N values $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ is given by:

$$\text{Var}(\{\boldsymbol{z}_i\}) = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N}\boldsymbol{z}_i^2 - N\bar{\boldsymbol{z}}^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^{N}\boldsymbol{z}_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N}\boldsymbol{z}_i\right)^2\right). \tag{26}$$

For $N = 0$ and $N = 1$, variance is typically undefined or taken as 0. We consider $N \geq 2$ for variance calculation. To compute $v_{\text{gen},[d]}$ incrementally for dimension d, when a new feature $\boldsymbol{z}_N$ (with d-th component $z_{N,[d]}$) is added, we need to maintain the sum of values and the sum of squared values for each dimension d of the generated features.

For each dimension $d \in \{1, \ldots, D\}$, the sum of feature values in dimension d for $N-1$ samples is:

$$\text{Sum}_d^{(N-1)} = \sum_{i=1}^{N-1} z_{i,[d]}. \tag{27}$$

The sum of squared feature values in dimension d for $N-1$ samples is:

$$\text{SumSq}_d^{(N-1)} = \sum_{i=1}^{N-1} (z_{i,[d]})^2. \tag{28}$$

(1) Initialization ($N = 0$)

Before any image is generated, for each dimension d:

$$\text{Sum}_d^{(0)} = 0, \tag{29}$$

$$\text{SumSq}_d^{(0)} = 0, \tag{30}$$

$$v_{\text{gen},[d]} = 0. \tag{31}$$

(2) First Sample ($N = 1$)

When the first feature $\boldsymbol{z}_1$ is generated:

$$\text{Sum}_d^{(1)} = z_{1,[d]}, \tag{32}$$

$$\text{SumSq}_d^{(1)} = (z_{1,[d]})^2, \tag{33}$$

$$v_{\text{gen},[d]} = 0. \tag{34}$$

(3) Subsequent Samples ($N > 1$)

Assume we have $\text{Sum}_d^{(N-1)}$ and $\text{SumSq}_d^{(N-1)}$ for $N-1$ samples, and a new feature $\boldsymbol{z}_N$ (with d-th component $z_{N,[d]}$) is generated:

$$\text{Sum}_d^{(N)} = \text{Sum}_d^{(N-1)} + z_{N,[d]}, \tag{35}$$

$$\text{SumSq}_d^{(N)} = \text{SumSq}_d^{(N-1)} + (z_{N,[d]})^2, \tag{36}$$

$$v_{\text{gen},[d]} = \frac{1}{N-1}\left(\text{SumSq}_d^{(N)} - \frac{1}{N}(\text{Sum}_d^{(N)})^2\right). \tag{37}$$

With $v_{\text{gen},[d]}$ calculated for all dimensions d using the updated sums, the $\mathcal{R}_{\text{VM}}$ can be computed.

E Discussion

E.1 Limitations

The proposed framework's performance is contingent upon access to high-quality, pre-trained diffusion models. Furthermore, the efficacy of DCS may be reduced when applied to highly specialized domains poorly represented in the diffusion model's training data, and the framework's multi-component reward system and iterative boosting loop introduce complexities in tuning and computational demand.

E.2 Broader impact

The DCS framework is a foundational research contribution aimed at improving the adaptability of machine learning models in data-scarce, continual learning settings. As the approach leverages a pre-trained diffusion model, the nature and characteristics of the generated images are directly determined by this underlying component. Consequently, the framework inherits any potential societal impacts, such as fairness considerations or biases, from the foundational model it employs. The safety and security of the generated content are therefore contingent on the specific diffusion model used, and we encourage practitioners to consider the ethical implications of the foundational model selected for any given application.

https://doi.org/10.52202/085713-2153

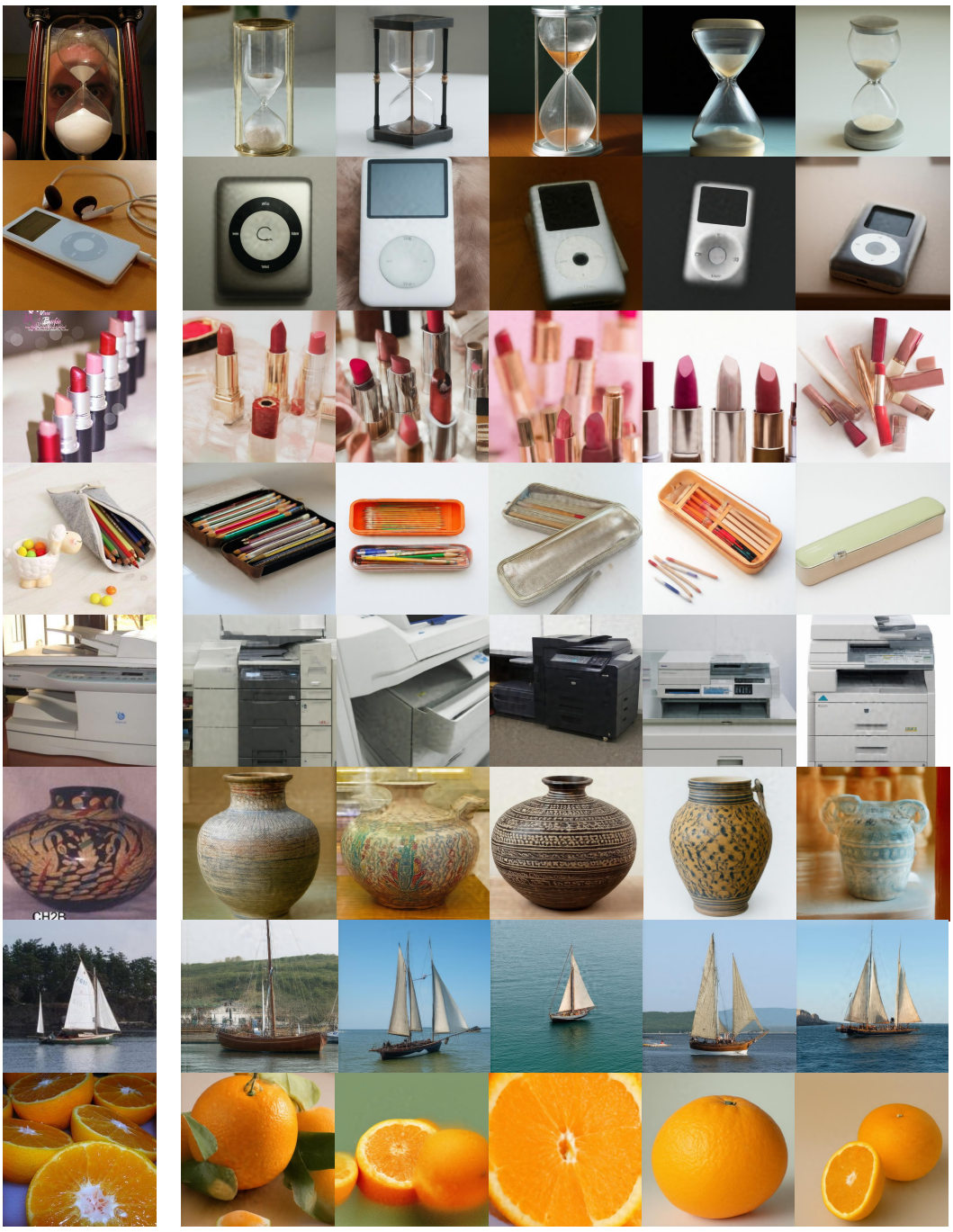

Figure 4: Qualitative comparison of real and generated images on *mini*ImageNet. The first column displays real images from the dataset. The subsequent columns show diverse and semantically correct images generated by our DCS framework for the corresponding classes.