Contrastive Conditional–Unconditional Alignment for Long-tailed Diffusion Model

Fang Chen[1], Alex Villa[1], Gongbo Liang[2], Li Fuxin[3], Xiaoyi Lu[4], and Meng Tang[1]

[1] University of California Merced, USA
[2] Texas A&M University-San Antonio, USA
[3] Oregon State University, USA
[4] University of Florida, USA

Abstract. Training data for class-conditional image synthesis often exhibit a long-tailed distribution with limited amount of images for tail classes. Such an imbalance causes mode collapse and reduces the diversity of synthesized images for tail classes. For class-conditional diffusion models trained with imbalanced data, we aim to improve the diversity and fidelity of tail class images without compromising the quality of head class images. We propose contrastive conditional-unconditional alignment (CCUA), which comprises two synergistic loss functions. Our first loss is an Alignment Loss (AL) that aligns class-conditional generation with unconditional generation at large timesteps. Alignment loss makes the denoising process insensitive to class conditions for the initial steps, which enriches tail classes through knowledge sharing from head classes. Secondly, we diversify unconditional generation via an Unsupervised Contrastive Loss (UCL) to increase the distance/dissimilarity among synthetic images. We combine the two losses to implicitly diversify conditional generation. Our framework is easy to implement as demonstrated on both U-Net based architecture and Diffusion Transformer. Our method outperforms vanilla denoising diffusion probabilistic models, score-based diffusion model, and alternative contrastive methods for class-imbalanced image generation across various datasets, in particular ImageNet-LT with 256×256 resolution.

1 Introduction

Recent advances in diffusion models [13,38] have led to breakthroughs in various generation tasks such as image generation [33,34,46], video generation [12,15], image editing [7,19], 3D generation [29], etc. These diffusion-based generative models rely on large-scale datasets for training, which often follow a long-tailed distribution with a significant amount of data for head classes and a limited amount of data for tail classes. Similar to the class-imbalanced recognition model [24] and the class-imbalanced generative adversarial network [20,31,32,40], diffusion model is often not able to generate high-quality images for tail classes due to the scarcity of training data. As shown in Fig. 1, the original diffusion model with transformer backbone [25] generates inferior images for a tail class

(a) SiT [25] (b) SiT with CCUA (ours)

Fig. 1: Generated images for a tail class ('red wine') with only 13 training images by (a) standard SiT [25] with diffusion loss and (b) proposed CCUA framework. Both models are trained on long-tailed ImageNet dataset for 900k steps with 256x256 resolution.

(red wine) when trained using a long-tailed version of ImageNet. We aim to increase the fidelity and diversity of tail class images while maintaining the quality of head class images.

We look into a highly related and crucial problem of long-tailed image recognition, for which contrastive learning is an effective method [16,23]. Rather than applying standard supervised contrastive learning [21] or unsupervised contrastive learning [27] for conditional generation, we propose to diversify unconditional generation via a repulsive force among samples and further align the conditional generation with unconditional generation via an attractive force. Fig. 2 illustrates how our contrastive conditional-unconditional alignment framework is different from other contrastive learning variants.

Specifically, our framework combines two synergistic losses. Our first loss is an unsupervised contrastive loss with negative samples only to diversify unconditional generation, serving as regularization for denoising diffusion probabilistic models (DDPM) and score-based diffusion models (SBDM). Given that mode collapse during inference manifests as generated samples being overly similar for tail classes, we introduce unsupervised contrastive learning with negative samples to maximize the distances among generated images. Unlike supervised contrastive learning [21], unsupervised contrastive learning distinguishes between representations of images regardless of their class, hence increasing intra-class and inter-class image diversity. Our unsupervised contrastive loss is implemented with batch resampling, which is a standard strategy for long-tailed recognition and generation [36,45,47].

Our second loss is an alignment loss designed to align estimated noises from conditional and unconditional generation, which effectively minimizes the KL divergence between latent distributions for conditional and unconditional generation. While such conditional-unconditional alignment seems undesirable with less controllability, a critical aspect is that our alignment loss is weighted more for larger timesteps corresponding to the initial stage of the reverse process.

Our alignment loss can also be motivated by the success of conditional-unconditional alignment for long-tailed GAN [20,35], which facilitates knowledge sharing between head classes and tail classes. Notably, unconditional GAN generation has been observed to achieve superior FID than conditional generation under limited data [35]. Unlike the GAN-based method [20] which aligns conditional generation and unconditional generation for low-resolution representations of images exhibiting intra-class similarity, we propose to match conditional generation with unconditional generation for large timesteps, by leveraging observed image similarity during the initial denoising steps for tail classes and head classes [37].

The synergy of the two losses makes our method effective, with the unsupervised contrastive loss diversifying *unconditional* latents repulsing negative pairs, and alignment loss aligning *conditional* and *unconditional* latents. As a result, conditional latents are diversified implicitly, which is theoretically and empirically better than directly diversifying conditional latents. Previous work utilized contrastive learning to improve adversarial robustness [28], find semantically meaningful directions [8], accelerate training [44], and regularize representation [41]. We effectively leveraged alignment and contrastive learning for class-imbalanced diffusion models and demonstrated the superior performance of our method on long-tailed image generation via comprehensive experiments.

The main contributions of this work are as follows:

- We propose **C**ontrastive **C**onditional-**U**nconditional **A**lignment (**CCUA**) for Diffusion Model with imbalanced data. Our proposed losses are easy to implement with DDPM and SBDM pipelines for both UNet-based architecture and Diffusion Transformer.
- Firstly, our Alignment Loss (AL) aligns unconditional generation and conditional generation for the initial steps in the denoising process, facilitating knowledge sharing between head and tail classes.
- Secondly, our Unsupervised Contrastive Loss (UCL) employs unsupervised contrastive learning with negative samples only, enhancing intra-class diversity for unconditional generation.
- We improved the diversity and fidelity of tail class for conditional generation while maintaining the quality of head class for multiple datasets and various resolutions, in particular ImageNet-LT with 256x256 resolution. Our framework outperforms other contrastive learning variants, such as the concurrent work of dispersive loss [41].

2 Related Work

Class-imbalanced Image Generation Generative models such as VAE [22], GAN [4,18], and diffusion models [13,33] generate inferior images for tail classes when trained with real-world data with a long-tailed distribution. It has attracted a lot of research interest [1,20,30,40,47] to address this issue for various types of models. Many methods address the problem of class imbalance by augmenting training data for the tail classes. A VAE is fine-tuned on tail classes under a majority-based prior [1]. It is observed that GAN [40] can amplify biases, leading to tail classes to be barely generated during inference, highlighting that the fairness in GAN needs improvement. Khorram et al. [20] propose a GAN-based long-tailed generation method, named UTLO, which shares the latent representations of conditional GAN with unconditional GAN and implicitly shares knowledge between the head class and tail class. The motivation with UTLO is that low-resolution representations of images from GAN are similar for head classes and tail classes. We observe a similar phenomenon for denoised images in the initial steps of the denoising process, and further propose a conditional-unconditional alignment loss designed for diffusion models. Recent work addresses class-imbalanced diffusion models [30, 43, 47] by regularization losses to align or separate the distributions of synthetic images and their corresponding latent representations across different classes. For example, CBDM [30] loss minimizes the distance of estimated noise items between these two models, the original DDPM model and a second model trained with pseudo labels which form a uniform distribution. DiffROP [43] attempts to combine contrastive learning with diffusion model by maximizing the distance of distributions between classes. However, DiffROP only considers pairs of images of different classes as negative pairs without regularizing images of the same class.

Contrastive Learning for Representation Learning Both unsupervised contrastive learning [6, 11] and supervised contrastive learning [16, 21, 23] are representation learning methods by maximizing the similarity between positive pairs and minimizing the similarity between negative pairs. Unsupervised contrastive learning does not require labels and typically augments the same data to form positive pairs [6,11]. Supervised contrastive learning incorporates class labels and pulls together all samples from the same class while pushing apart samples from different classes. We propose a new variant of contrastive learning with conditional-unconditional alignment for class-imbalanced diffusion models. Our work is different from the contrastive-guided diffusion process [28], which aims to improve adversarial robustness. While the diffusion model originally for generation tasks becomes an emerging technique for representation learning [2,10], we directly integrate contrastive learning regularization to diffusion models for better generation on imbalanced data. Notably, unsupervised contrastive loss with negative samples exclusively focuses on separating dissimilar instances in the embedding space, which we leverage to address mode collapse, as detailed in Sec. 3.2. In the context of generative modeling, REPA [44] aligns diffusion model features with features from a frozen vision encoder to accelerate training. Unlike REPA, we directly regularize diffusion model features without external models. Most relevant is concurrent work of dispersive loss [41], which is similar to our unsupervised contrastive loss with negative pairs only. Dispersive loss is directly and explicitly applied to conditional training, as designed for balanced diffusion models. By comparison, our unsupervised contrastive loss is formulated in the unconditional latent space and is implicitly extended to conditional training through our conditional–unconditional alignment loss. This carefully designed contrastive conditional-unconditional alignment framework for long-tailed diffusion models achieves better generation as shown in experiments.

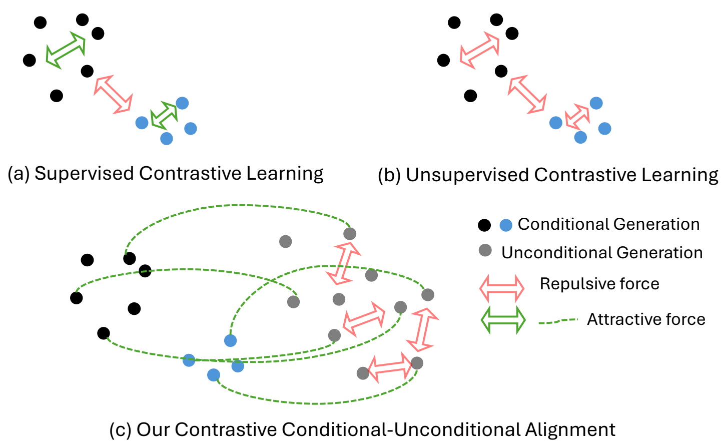

Fig. 2: Variants of contrastive learning for regularizing latents of conditional & unconditional generation and our proposed CCUA framework. The black and blue dots are latents from class-conditional generation and grey dots are from unconditional generation. The latents represent intermediate features from a U-Net or transformer-based denoising network. Rather than applying standard supervised contrastive learning (a) and unsupervised contrastive learning (b) for *conditional* generation, we propose to diversity *unconditional* generation via repulsive force and further align *conditional* generation with *unconditional* generation via attractive force (c).

3 Method

We propose **C**ontrastive **C**onditional-**U**nconditional **A**lignment (CCUA) for Diffusion Model. We show the intuition and motivation in Sec. 3.1. Our method involves a unsupervised contrastive loss for unconditional generation and a conditional-unconditional alignment loss, as detailed in Sec. 3.2 and 3.3. Sec. 3.4 summarizes our framework and discusses the synergy between the two losses.

3.1 Motivation of CCUA Framework

Limited amount of training images for tail classes leads to limited diversity for synthetic images. Batch sampling with replacement leads to more severe overfitting on training tail class samples. We explore contrastive learning, which is shown effective for long-tailed recognition [16, 21, 23], for the new task of long-tailed generation. The idea is to regularize latents of a denoising network at timestep t, which can be intermediate feature or output noise from a U-Net or a transformer. We discuss variants of contrastive learning as regularization for latents, and how our proposed framework is different.

Supervised contrastive learning [21] as shown in Fig. 2 (a) employs an attractive force for latents of the same class and a repulsive force for latents of different classes. However, supervised contrastive learning doesn't improve **intra-class**

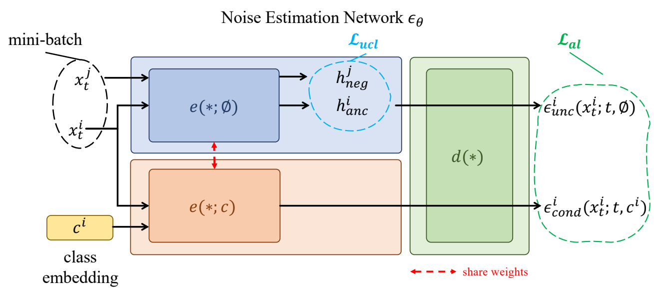

Fig. 3: Implementation of the proposed CCUA framework. The noise estimation network is divided into a latent encode network $e(*)$ and a decode network $d(*)$. $e(*)$ encodes input x_t to a low-dimensional latent h, which is decoded to noise ϵ by $d(*)$. We increase the distance of unconditional latents by $\mathcal{L}_{ucl}$ with negative samples only, and align unconditional and conditional generation from the same sample x_t^i and utilize $\mathcal{L}_{al}$ to minimize their distances at initial time steps. Specifically, $e(*)$ and $d(*)$ for unet-based model and diffusion transformer are shown in Fig. 6 in Appendix A.1.

diversity as it involves an attractive force for pairs of images of the same class. A better way to improve diversity is to have repulsive force for all intra-class images and inter-class images, which can be implemented via InfoNCE loss [27] with negative samples only in the concurrent work of dispersive loss [41].

We reveal the theoretical limitations of such unsupervised contrastive learning on conditional generation in Sec. 3.2. We propose a dramatically different method that diversifies unconditional latents and also aligns conditional latents with unconditional latents shown in Fig. 2 (c). A repulsive force on unconditional latents improves diversity in particular for tail classes, and an attractive force aligns conditional generation with unconditional generation for large timesteps. The two losses combined implicitly diversify latents from conditional generation. Fig. 3 gives an overview of how our framework is implemented.

3.2 Unsupervised Contrastive Loss for Unconditional Generation

We observed the synthesized images of DDPM for a tail class concentrated around the limited training images in the latent space, leading to mode collapse, as shown in Fig. 7, and Fig. 10 in Appendix A.5.

We consider contrastive loss with negative samples only, where the negative samples are based on noisy images from a mini-batch. We treat a synthetic image and itself as a positive pair and all other images within the batch as negative samples. By applying the contrastive loss with negative samples only, we increase the distance of each synthesized image from other images in the latent space, hence increasing the diversity of synthetic images in particular for tail classes. Specifically, we divide a noise estimation network into two parts, a latent encode network $e(*)$, and decode network $d(*)$. $e(*)$ encodes an image with noise x_t to a low-dimensional latent h, which is decoded to the estimate noise item ϵ by $d(*)$. Our contrastive loss is defined for the latents h with each image in a mini-batch treated as an anchor. All other images are treated as negative samples. The anchor sample x_t^i and negative samples x_t^j are fed into the encoder $e(*)$ to get unconditional latents $h_{anc}^i = e_\theta(x_t^i; t, \emptyset)$ and $h_{neg}^j = e_\theta(x_t^j; t, \emptyset)$. Our unsupervised contrastive loss $\mathcal{L}_{ucl}$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{ucl} = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\pi_{anc}^i}{\pi_{anc}^i + \sum_{j\in B, j\neq i}\pi_{neg}^j}, \\ \text{with } \pi_{anc}^i = \exp(\frac{h_{anc}^i \cdot h_{pos}^i}{\tau}) = \exp(\frac{1}{\tau}), \pi_{neg}^j = \exp(\frac{h_{anc}^i \cdot h_{neg}^j}{\tau}), \end{aligned} \tag{1}$$

where B denotes a mini-batch, and τ is a temperature for softmax which we keep 0.1 as the default setting. Note that we do not augment an anchor image x_t^i to be a positive sample like many unsupervised contrastive learning methods. Instead, we directly treat the anchor latent h_{anc}^i itself as the positive vector in the contrastive loss, i.e., $h_{pos}^i := h_{anc}^i$. We also normalized latents following standard protocol for InfoNCE loss [27]. We discuss implementation details on batch sampling for our UCL loss in Appendix A.1.

Why unsupervised contrastive loss improves image diversity? In the worst case, that all embeddings collapse to a constant vector, the contrastive loss becomes $\log|B|$ for a batch with $|B|$ samples. Such constant embeddings yield the maximum possible loss and are therefore discouraged. Intuitively, contrastive loss introduces a repulsive force between different samples. The numerator remains constant, as self-similarity is always maximal, while the denominator aggregates pairwise similarities across the batch. Minimizing the loss requires reducing similarities between distinct samples, effectively pushing their embeddings apart in feature space. Geometrically, the optimal solution corresponds to embeddings being uniformly distributed on a hypersphere.

Limitation of diversifying conditional latents directly It is known that InfoNCE loss [27] for contrastive learning gives a lower bound of mutual information between data sample and representation/latents. In other words, minimizing InfoNCE loss maximizes mutual information. We denote conditional latent as $h_t^c := e_\theta(x_t, t, c)$ and unconditional latent as $h_t^u := e_\theta(x_t, t, \emptyset)$, where c is the class condition. The InfoNCE loss on unconditional latents maximizes the mutual information $I(h_t^u; x)$ between image x and its unconditional latent h_t^u. Similarly, InfoNCE loss on conditional latents [41] maximizes the mutual information $I(h_t^c; x, c)$ between image x and its conditional latent h_t^c and condition c.

Based on the chain rule of mutual information, $I(h_t^c; x, c)$ can be decomposed into two parts:

$$I(h_t^c; x, c) = I(h_t^c; c) + I(h_t^c; x|c). \tag{2}$$

The first item $I(h_t^c; c)$ measures the mutual information between class c and latent h_t^c, while the second item $I(h_t^c; x|c)$ measures the conditional mutual information. Intuitively, the first term tells how much class condition reveals about the latent, which reflects **inter-class diversity**. The second term measures conditional mutual information with condition c which reflects **intra-class diversity**. To address the mode collapse issue for tail classes and increase diversity, it is apparent that we need to focus on maximizing the second item, i.e., conditional mutual information $I(h_t^c; x|c)$. However, the optimization of inter-class mutual information $I(h_t^c; c)$ admits a trivial solution of having an identical latent for all images of the same class leading to maximum mutual information.

Unconditional Contrastive Loss and Combination with Alignment Loss There is another alternative decomposition:

$$I(h_t^c; x, c) = I(h_t^c; x) + I(h_t^c; c \mid x), \tag{3}$$

where $I(h_t^c; x)$ dominates since the $I(h_t^c; c|x)$ varies little. $I(h_t^c; x)$ adopts a lower bound $I(h_t^c; x) \geq I(h_t^u; x) + I(h_t^c; h_t^u) - H(h_t^u)$, where h_t^u is the unconditional latent. Our unconditional contrastive loss on unconditional latents increases $I(h_t^u; x)$ by pushing different samples apart regardless of class labels. Our alignment loss in Sec. 3.3 reduces the discrepancy between h_t^c and h_t^u, which increases $I(h_t^c; h_t^u)$. Thus, the two losses jointly increase a lower bound on $I(h_t^c; x)$, encouraging the conditional latent to preserve more sample-level variation.

3.3 Conditional-unconditional Alignment Loss

As shown in Fig. 2, one of our goals is to align the distribution of latents from conditional and unconditional generation, which implicitly diversify conditional generation. Specifically, we penalize KL divergence between conditional distribution $p_\theta(x_{t-1}|x_t^i, c^i)$ and unconditional distribution $p_\theta(x_{t-1}|x_t^i)$:

$$\mathcal{L}_{al}^{i,t} = D_{KL}[p_\theta(x_{t-1}|x_t^i, c^i)||p_\theta(x_{t-1}|x_t^i)]. \tag{4}$$

Suppose

$$\begin{aligned} p_\theta(x_{t-1}|x_t, c) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \sigma_t^2 I), \\ p_\theta(x_{t-1}|x_t) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I), \end{aligned} \tag{5}$$

then we have:

$$\mathcal{L}_{al}^{i,t} = \mathbb{E}[\frac{1}{2\sigma_t^2}||\mu_\theta(x_t^i, t, c^i) - \mu_\theta(x_t^i, t)||^2] + C, \tag{6}$$

where C is constant, and

$$\begin{aligned} \mu_\theta(x_t^i, t, c^i) &= \frac{1}{\sqrt{\alpha_t}} x_t^i - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t^i, t, c^i), \\ \mu_\theta(x_t^i, t) &= \frac{1}{\sqrt{\alpha_t}} x_t^i - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t^i, t), \end{aligned} \tag{7}$$

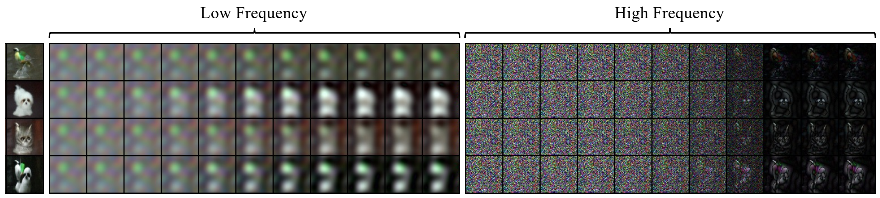

Fig. 4: The leftmost column shows synthetic images with different classes but with the same initial noise or random seed. We visualize the low-frequency component and the high-frequency component from the reverse processing. It shows that low-frequency components are similar during initial time steps for different classes with the same initial noise [37]. More examples are in Appendix A.5.

where $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, $\{\beta_t\}_{1:T}$ is the variance schedule. Then, $\mathcal{L}_{al}$ simplifies to:

$$\mathcal{L}_{al}^{i,t} \propto ||\epsilon_\theta(x_t^i, t, c^i) - \epsilon_\theta(x_t^i, t)||^2. \tag{8}$$

As shown in Fig. 3, the anchor vector h_{anc}^i is fed into decode network $d(*)$ to get unconditional noise estimation $\epsilon_{unc}^i := \epsilon_\theta(x_t^i; t, \emptyset)$. Meanwhile, the anchor image x_t^i is also incorporated with the class label condition $\mathbf{c^i}$ fed into the network to get the conditional noise estimation $\epsilon_{cond}^i := \epsilon_\theta(x_t^i; t, \mathbf{c^i})$. The alignment loss is defined between the unconditional and conditional noise estimation:

$$\mathcal{L}_{al} = \frac{1}{|B|}\sum_{i\in B}\mathbb{E}_{t,x_0}[\frac{t}{T}\|\epsilon_\theta(x_t^i; t, \mathbf{c^i}) - \epsilon_\theta(x_t^i; t, \emptyset)\|^2]. \tag{9}$$

The loss is weighted linearly by timesteps t, so that the initial steps with large t are weighted more. In other words, we align conditional generation and unconditional generation for the initial steps.

Connection to Conditional-unconditional Alignment for Long-tailed GAN We take inspiration from UTLO [20] and Transitional-GAN [35], which addresses long-tailed generation with GANs, and observes that the similarity between head class and tail class images increases at lower resolution representations. To share knowledge between the head class and tail class, UTLO [20] proposes unconditional GAN objectives for lower-resolution representations and conditional GAN objectives for subsequent higher-resolution images. More specifically, utilizing unconditional generation for lower resolution is effective in increasing the diversity and quality of tail class images.

Our alignment loss for diffusion model is similar to the conditional-unconditional alignment approach used in GANs [20,35]. Our key insight is that diffusion models denoise images recursively, and the similarity between head class images and tail class images is higher during the *initial timesteps* of the denoising process. This is similar but different from UTLO [20], which proposes unconditional generation for *lower-resolution* representations. We visualize the reverse processing of unconditional generation and conditional generation with different class labels of the original DDPM, starting from the same Gaussian noise (Fig. 4).

Images from unconditional generation and conditional generation share similar low-frequency components. Our alignment loss matches unconditional generation and conditional generation at the initial time steps, enabling knowledge sharing between tail classes and head classes with abundant data.

3.4 Overall Framework

Our final loss function $\mathcal{L}$ is the sum of the standard DDPM [13] loss $\mathcal{L}_{ddpm}$ and our contrastive conditional-unconditional alignment loss $\mathcal{L}_{ccua}$:

$$\mathcal{L}_{ccua} = \alpha \cdot \mathcal{L}_{ucl} + \gamma \cdot \mathcal{L}_{al}, \tag{10}$$

where α and γ are the hyper-parameters for our unsupervised contrastive loss and alignment loss. The overall algorithm framework is shown as Algorithm 1. The parameters of the noise estimation network ϵ_θ are updated by the gradient of the final loss $\nabla_\theta \mathcal{L}$ as in Eq. 10.

Algorithm 1 Training algorithm of **CCUA**.

Set $\mathcal{L}_{ddpm}, \mathcal{L}_{ucl}, \mathcal{L}_{al} = 0$
for each image-class pair (x_0^i, c^i) in this batch B **do**
  Sample $\epsilon^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \mathcal{U}(\{0, 1, ..., T\})$
  $x_t^i = \sqrt{\bar{\alpha}_t} x_0^i + \sqrt{1 - \bar{\alpha}_t} \epsilon^i$
  $\epsilon_{unc}^i = \epsilon_\theta(x_t^i; t, \emptyset)$,
  $\epsilon_{cond}^i = \epsilon_\theta(x_t^i; t, c^i)$,
  $h_{anc}^i = e(x_t^i)$,
  $\pi_{anc} = \exp(h_{anc}^i \cdot h_{anc}^i / \tau)$,
  Set $\pi_{neg} = 0$
  for x_t^j in this batch with $i \neq j$ **do**
    $h_{neg}^j = e(x_t^j)$
    $\pi_{neg} = \pi_{neg} + \exp(h_{anc}^i \cdot h_{neg}^j / \tau)$
  end for
  $\mathcal{L}_{ucl} = \mathcal{L}_{ucl} + (-\log \frac{\pi_{anc}}{\pi_{anc} + \pi_{neg}})$
  if Unconditional Training of CFG **then**
    $\mathcal{L}_{ddpm} = \mathcal{L}_{ddpm} + \|\epsilon^i - \epsilon_{unc}^i\|^2$
  else if Conditional Training of CFG **then**
    $\mathcal{L}_{ddpm} = \mathcal{L}_{ddpm} + \|\epsilon^i - \epsilon_{cond}^i\|^2$
    $\mathcal{L}_{al} = \mathcal{L}_{al} + \frac{t}{T} \|\epsilon_{unc}^i - \epsilon_{cond}^i\|^2$
  end if
end for
$\mathcal{L}_{ccua} = \frac{1}{|B|}(\mathcal{L}_{ddpm} + \alpha \cdot \mathcal{L}_{ucl} + \gamma \cdot \mathcal{L}_{al})$

Synergy between unsupervised contrastive loss and alignment loss Our proposed two losses including unsupervised contrastive loss $\mathcal{L}_{ucl}$ and $\mathcal{L}_{al}$ are novel by themselves in the context of diffusion models. What's more, the synergy of the two losses makes our method effective. Our $\mathcal{L}_{ucl}$ loss diversifies *unconditional* latents repulsing negative pairs, while $\mathcal{L}_{al}$ aligns *conditional* and *unconditional* latents/representations. As a result, conditional latents are diversified implicitly. We choose not to directly diversify conditional latents, as is done in concurrent work of Dispersive Loss [41]. Our combined loss is more effective for addressing overfitting in tail classes empirically verified in Tab. 1. The geometric interpretation for the reason is that the diversity of conditional latents involves both inter-class variance and intra-class variance. Directly applying InfoNCE loss on conditional latent [41] can lead to a shortcut solution that dominantly maximizes inter-class variance. However, contrastive regularization on unconditional latents *must* diversity latents for all data regardless of class. Aligning conditional latents to diversified unconditional latents via distillation is more effective.

Table 1: Comparison on ImageNet-LT 256 × 256 with SiT pipeline. We use blue parentheses '()' to highlight the improvement of our method over SiT baseline on FID and Recall, which denotes the overall quality and diversity, respectively. Our method dramatically improves FID$_{tail}$ in particular.

Steps	Method	IS ↑	FID ↓	sFID ↓	Prec. % ↑	Recall % ↑	FID$_{tail}$ ↓
250k	SiT	53.9	33.8	22.6	54.5	19.1	52.8
	CBDM	54.8	34.1	23.3	53.9	18.7	53.1
	REPA	**74.1**	28.4	**20.0**	58.3	16.1	48.5
	Dispersive Loss	53.8	34.0	22.6	54.8	19.8	54.5
	CCUA (ours)	70.7	**27.0** (-6.8)	20.5	**60.2**	**20.1**	**37.0** (-15.8)
450k	SiT	78.8	25.7	21.9	64.3	18.5	41.6
	CBDM	84.0	24.7	21.8	64.6	18.7	40.7
	REPA	105.9	21.9	19.5	65.7	15.2	39.7
	Dispersive Loss	83.8	25.2	21.7	65.5	17.3	43.0
	CCUA (ours)	**111.9**	**19.4** (-6.3)	**18.3**	**69.4**	**18.9**	**28.8** (-12.8)
700k	SiT	103.1	21.2	20.1	69.3	18.2	35.4
	CBDM	105.7	20.94	20.7	70.3	17.8	35.5
	REPA	126.8	19.7	20.3	68.0	15.8	36.9
	Dispersive Loss	104.0	21.3	20.4	68.0	**18.5**	35.9
	CCUA (ours)[5]	**140.5**	**16.3** (-4.9)	**17.2**	**73.9**	18.4	**24.5** (-10.9)
900k	SiT	111.7	19.9	20.1	70.3	18.6	33.9
	CBDM	117.2	19.4	20.2	72.5	17.7	32.7
	REPA	137.8	18.1	19.3	69.8	16.2	33.8
	Dispersive Loss	115.6	19.7	20.1	69.9	**18.9**	34.1
	CCUA (ours)	**153.1**	**15.1** (-4.8)	**16.5**	**75.7**	17.3	**22.5** (-11.4)

4 Experiments

4.1 Experimental Setup

We report main results on ImageNet, TinyImageNet, and Places datasets, with their long-tailed version, while we also report results on CIFAR10/CIFAR100 long-tailed datasets as shown in Appendix A.3. We measure IS, FID, KID [3], spatial FID [9] (sFID), Precision, Recall, and FID of tail classes (FID$_{tail}$) as the evaluation metrics. More implementation details are provided in Appendix A.1.

4.2 Quantitative Results

Class-imbalanced Generation for Diffusion Transformer We apply the proposed CCUA framework into SiT [25] on ImageNet-LT dataset, and compare to CBDM [30], a long-tailed training method for UNet-based model, and REPA [44],

[5] Our method incurs longer training time as discussed in Appendix. A.2. However, CCUA still outperforms SiT under the same training time (i.e., 700k for CCUA, 900k for SiT).

Table 2: Comparison on TinyImageNet-LT 64×64 and Places-LT 64×64 with DDPM pipeline. Blue '()' shows improvement of our method over DDPM baseline. Green '()' shows improvement of DDPM trained on balanced version over long-tailed version.

Dataset	Method	FID↓	FID_{tail} ↓	$\text{KID}_{\times 1k}$ ↓
TinyImageNet-LT	DDPM^*_{bal} [13]	15.7 (-3.0)	25.6 (-14.5)	3.2 (-3.1)
	DDPM [13]	18.7	40.1	6.3
	CBDM [30]	20.9	48.1	6.6
	OCLT [47]	17.7	39.7	5.6
	CCUA (ours)	**15.2** (-3.5)	**30.4** (-9.7)	**3.8** (-2.5)
Places-LT	DDPM [13]	13.9	23.7	5.3
	CBDM [30]	15.2	26.1	5.6
	OCLT [47]	13.0	22.8	4.2
	CCUA (ours)	**12.0** (-1.9)	**20.8** (-2.9)	**3.6** (-1.7)

the latest training technique for general DiT/SiT-based model. We also compare our method to the concurrent work of Dispersive Loss [41] which is a contrastive loss originally developed for improving diffusion models on balanced datasets. As shown in Tab. 1, our method achieves remarkable improvement compared to SiT, CBDM, REPA and Dispersive Loss on various training steps. Our method achieves about 20% improvement on overall FID, 30% improvement on IS Score and FID_{tail}, illustrating the effectiveness of CCUA.

Class-imbalanced Generation for DDPM For training unet-based architecture DDPM on TinyImageNet-LT and Places-LT datasets, our method also achieves the best performance, as shown in Tab. 2. For TinyImageNet-LT datasets, we provide the metrics of the DDPM model trained on the original balanced datasets, denoted by DDPM^*_{bal}, as the theoretical optimal reference. The best FID and KID score of our method illustrates the high quality of images synthesized by our method.

We provide a comparison to other methods for long-tailed diffusion models including DiffROP [43] in Appendix A.3.

Results Across Various Category Intervals To see the effectiveness of our method on tail classes, we divide each dataset into three super categories: 'Head', 'Body', and 'Tail', with classes sorted by the number of training images. For each dataset, the top 33% classes were allocated to the 'Head' category, the next 34% classes to the 'Body' category, and the rest to the 'Tail' category. The percentage $P_{category}$ of training images belonging to each category is shown in the header of Tab. 3. Our method achieves better FID score for 'Body' and 'Tail' categories, while keeps 'Head' category unchanged or even better.

4.3 Qualitative Results and Ablation Study

Visualization of Synthetic Images Fig. 5 shows synthetic images of SiT and with CCUA for ImageNet-LT tail classes 'espresso' and 'window shade'.

Table 3: FID score for three 'super-categories': 'Head', 'Body', and 'Tail'.

Dataset	FID↓ $P_{category}$ / Method	Head $\sim 80\%$	Body $\sim 17\%$	Tail $\sim 3\%$	All
TinyImageNet-LT	DDPM [13]	21.3	33.3	40.1	18.7
	CBDM [30]	24.1	35.0	48.1	20.9
	OCLT [47]	**20.6**	30.6	39.7	17.7
	CCUA (ours)	21.3	**28.0**	**30.4**	**15.2**
Places-LT	DDPM [13]	19.3	20.3	23.7	13.9
	CBDM [30]	21.3	21.7	26.1	15.2
	OCLT [47]	19.2	19.4	22.8	13.0
	CCUA (ours)	**18.2**	**19.3**	**20.8**	**12.0**

(a) Synthetic images from SiT [25].

(b) Synthetic images from CCUA (ours).

Fig. 5: Synthetic images of SiT and it with CCUA for ImageNet-LT tail classes 'espresso' and 'window shade'. All methods start the denoising process from the same Gaussian noise at corresponding grid cells. CCUA shows consistently better diversity and fidelity.

Compared to SiT, CCUA achieves visually higher fidelity and diversity. More qualitative results on ImageNet-LT, TinyImageNet-LT, and CIFAR100-LT are shown in Appendix A.5.

Ablation Study of Batch Resample Strategy We conduct ablation study of batch resample strategy on both TinyImageNet-LT and ImageNet-LT datasets. As shown in Tab. 4, applying $\mathcal{L}_{ccua}$ itself only is already comparable to DDPM with batch resampling replacement. While with applying batch resample strategy, the proposed $\mathcal{L}_{ccua}$ loss benefits more from it and outperforms 'DDPM + Batch Resample' on all three metrics. The ablation study of Batch Resample Strategy on ImageNet-LT is reported in Tab. 12 in Appendix. A.4.

Ablation Study of Hyper-parameters and Losses Tab. 5 shows the ablation study of the hyper-parameters for the proposed $\mathcal{L}_{ucl}$ and $\mathcal{L}_{al}$ losses,

Table 4: Batch Resample Strategy Analysis on TinyImageNet-LT. Blue '()' highlights the improvement of each method compared to the DDPM baseline.

Method	FID ↓	FID_{tail} ↓	KID↓
DDPM (baseline)	18.7	40.1	6.3
+Batch Resample	17.2 (-1.5)	33.6 (-6.5)	4.7 (-1.6)
$\mathcal{L}_{ccua}$ (Eq. 10)	17.2 (-1.5)	37.9 (-2.2)	4.9 (-1.4)
+Batch Resample	**15.2** (-3.5)	**30.4** (-9.7)	**3.8** (-2.5)

Table 5: Hyper-parameters Analysis of CCUA on ImageNet-LT 256 × 256 with DiT-based models. All models are trained from scratch to 250k steps.

Method	α	γ	Latent	IS ↑	FID ↓	sFID ↓	Prec. (%) ↑	Recall (%) ↑
SiT (baseline)	0	0	-	53.9	33.8	22.6	54.5	19.1
$\mathcal{L}_{ccua}$ **(ours)**	1.0	1.0	N/4	58.2	30.5	**16.2**	51.1	**29.4**
	0.1	0.1	N/4	70.3	27.3	18.8	59.2	21.3
	0.1	0	N/4	67.40	28.76	25.48	58.52	19.30
	0	0.1	N/4	64.49	29.63	24.81	56.40	20.55
	0.05	0.05	N/4	**70.7**	**27.0**	20.5	60.2	20.1
	0.05	0	N/4	69.2	27.7	19.9	59.4	20.6
	0	0.05	N/4	68.4	28.2	19.5	58.8	21.6
	0.01	0.01	N/4	69.7	27.6	19.3	**60.7**	19.9
	1.0	1.0	N	52.5	33.0	**16.3**	48.9	**29.9**
	0.1	0.1	N	65.3	29.5	25.5	56.3	20.6
	0.05	0.05	N	**73.6**	**25.9**	16.5	58.6	23.1
	0.05	0	N	64.3	28.7	18.7	57.5	21.3
	0	0.05	N	64.7	29.0	20.2	**58.6**	20.0
	0.01	0.01	N	65.8	28.6	19.0	58.5	19.8

which is conducted on ImageNet-LT dataset. The 'Latent' denotes the latent representation h used for $\mathcal{L}_{ucl}$ is from the $N/4$-th or N-th block for a SiT model with N SiT/DiT blocks. We selected the optimal $\alpha = 0.05$, $\gamma = 0.05$ for the best FID and IS. We provide more detailed ablation study about transformer block applied to $\mathcal{L}_{ccua}$ shown in Tab. 13 in Appendix. A.4.

4.4 Limitation of our method

Our MSE loss involves both conditional generation and unconditional generation, which requires two passes of the denoising network during training and increases training time. With optimized implementation, the training time of our method is 1.3x of that of SiT, see Appendix A.2 for details. However, our method doesn't increase inference latency.

5 Conclusion

Real-world data for training image generation models often exhibit long-tailed distributions. Similar to class-imbalanced GANs [20], class-imbalanced diffusion models generates inferior tail class images due to data scarcity. We propose a framework with two losses that are synergistic. Firstly, we align class-conditional generation with unconditional generation for large timesteps using an alignment loss. This encourages the initial denoising steps to be class-agnostic, thereby enriching tail classes through knowledge sharing from head classes–a principle demonstrated to enhance long-tailed GAN performance [20,35], which we successfully adapt to diffusion models. Secondly, we diversify unconditional generation via an unsupervised contrastive loss with negative samples only to contrast the latents of different synthetic images, promoting intra-class diversity in particular for tail classes. With the two losses, our framework of contrastive conditional-unconditional alignment boosts the performance of DDPM [13] and SiT [25] for long-tailed image generation and outperforms alternative methods for long-tailed generation including CBDM [30], OCLT [47], and concurrent work including Dispersive Loss [41]. Extensive experiments on multiple datasets in particular ImageNet-LT with 256x256 resolution demonstrated the effectiveness of our method on both UNet-based architecture and Diffusion Transformer.

Acknowledgement

Research was sponsored by CAHSI-Google Institutional Research Program and by the Office of Naval Research under Grant Number W911NF-24-1-0295.

References

1. Ai, Q., Wang, P., He, L., Wen, L., Pan, L., Xu, Z.: Generative oversampling for imbalanced data via majority-guided vae. In: International Conference on Artificial Intelligence and Statistics. pp. 3315–3330. PMLR (2023)
2. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019)
5. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
7. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)

8. Dalva, Y., Yanardag, P.: Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24209–24218 (2024)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Fuest, M., Ma, P., Gui, M., Fischer, J.S., Hu, V.T., Ommer, B.: Diffusion models and representation learning: A survey. arXiv preprint arXiv:2407.00783 (2024)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
12. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)
16. Jiang, Z., Chen, T., Mortazavi, B.J., Wang, Z.: Self-damaging contrastive learning. In: ICML (2021)
17. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
19. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
20. Khorram, S., Jiang, M., Shahbazi, M., Danesh, M.H., Fuxin, L.: Taming the tail in class-conditional gans: Knowledge sharing via unconditional training at lower resolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7580–7590 (2024)
21. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
22. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019)
23. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R.S., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6918–6928 (2022)
24. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Stella, X.Y.: Open long-tailed recognition in a dynamic world. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
25. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)

26. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. Ouyang, Y., Xie, L., Cheng, G.: Improving adversarial robustness through the contrastive-guided diffusion process. In: International Conference on Machine Learning. pp. 26699–26723. PMLR (2023)
29. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)
30. Qin, Y., Zheng, H., Yao, J., Zhou, M., Zhang, Y.: Class-balancing diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18434–18443 (2023)
31. Rangwani, H., Bansal, L., Sharma, K., Karmali, T., Jampani, V., Babu, R.V.: Noisytwins: Class-consistent and diverse image generation through stylegans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5987–5996 (2023)
32. Rangwani, H., Jaswani, N., Karmali, T., Jampani, V., Babu, R.V.: Improving gans for long-tailed data through group spectral regularization. In: European Conference on Computer Vision. pp. 426–442. Springer (2022)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
35. Shahbazi, M., Danelljan, M., Paudel, D.P., Gool, L.V.: Collapse by conditioning: Training class-conditional GANs with limited data. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=7TZeCsNOUB_
36. Shi, J.X., Wei, T., Xiang, Y., Li, Y.F.: How re-sampling helps for long-tail learning? Advances in Neural Information Processing Systems **36**, 75669–75687 (2023)
37. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4733–4743 (2024)
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
40. Tan, S., Shen, Y., Zhou, B.: Improving the fairness of deep generative models without retraining. arXiv preprint arXiv:2012.04842 (2020)
41. Wang, R., He, K.: Diffuse and disperse: Image generation with representation regularization. arXiv preprint arXiv:2506.09027 (July 2025)
42. Wang, X., Dufour, N., Andreou, N., Cani, M.P., Abrevaya, V.F., Picard, D., Kalogeiton, V.: Analysis of classifier-free guidance weight schedulers. arXiv preprint arXiv:2404.13040 (2024)
43. Yan, D., Qi, L., Hu, V.T., Yang, M.H., Tang, M.: Training class-imbalanced diffusion model via overlap optimization. arXiv preprint arXiv:2402.10821 (2024)

44. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
45. Zhang, J., Liu, L., Wang, P., Shen, C.: To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions. arXiv preprint arXiv:1912.04486 (2019)
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
47. Zhang, T., Zheng, H., Yao, J., Wang, X., Zhou, M., Zhang, Y., Wang, Y.: Long-tailed diffusion models with oriented calibration. In: The Twelfth International Conference on Learning Representations (2024)