Reweighted Flow Matching via Unbalanced Optimal Transport for Label-free Long-tailed Generation

Hyunsoo Song
National Institute for Mathematical Sciences
song@nims.re.kr

Minjung Gim*
National Institute for Mathematical Sciences
mjgim@nims.re.kr

Jaewoong Choi*
Sungkyunkwan University
jaewoongchoi@skku.edu

Abstract

Flow matching has recently emerged as a powerful framework for continuous-time generative modeling. However, when applied to long-tailed distributions, standard flow matching frameworks are susceptible to *majority bias*, over-representing dominant modes at the expense of tail-class fidelity and distributional accuracy. In this work, we propose ***Unbalanced Optimal Transport Reweighted Flow Matching (UOT-RFM)***, a novel framework for generative modeling under class-imbalanced (long-tailed) distributions that operates without any class label information. Our method constructs the conditional vector field using mini-batch Unbalanced Optimal Transport (UOT) and mitigates majority bias through a principled inverse reweighting strategy. The reweighting relies on a *label-free majority score*, defined as the density ratio between the target distribution and the UOT marginal. This score quantifies the degree of majority based on the geometric structure of the data, without requiring class labels. By incorporating this score into the training objective, UOT-RFM theoretically recovers the target distribution with first-order correction ($k = 1$) and empirically improves tail-class generation through higher-order corrections ($k > 1$). Our model outperforms existing flow matching baselines on long-tailed benchmarks, while maintaining competitive performance on balanced datasets.

1 Introduction

Generative modeling aims to learn a model that can approximate complex target data distributions. In recent years, deep generative models have achieved remarkable progress across various domains, such as GANs [1, 16], optimal transport maps [7, 10, 35], and diffusion models [18, 37]. Among these, flow matching models [27] have emerged as a promising approach for continuous-time generative modeling. Flow matching learns a Continuous Normalizing Flow [4], which is a vector field that transports samples from a source distribution to a target distribution. The flow matching model is trained through regression to a conditional vector field, constructed from the conditional probability path between source and target samples, thereby avoiding the heavy computational costs associated with numerical likelihood estimation.

Despite these computational advantages, flow matching models face similar challenges as other generative approaches when dealing with real-world data characteristics. A particularly challenging scenario is the **long-tailed or imbalanced distributions**, where a few majority (head) classes

*Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

dominate and many minority (tail) classes are severely underrepresented [44]. Standard generative models often suffer from **majority bias** under such settings: head classes are well-modeled, but tail classes are badly generated or ignored [3, 32]. This leads to inaccurate class proportions, reduced sample diversity, and degraded generation quality for tail classes. To mitigate this, various long-tailed generative models have been proposed, such as GAN-based approaches [33, 34] and diffusion-based methods [32, 45]. While effective, these methods rely heavily on **explicit class label information** to improve generation for minority classes. Despite these advances, the **flow matching framework has not yet been explored** in the context of long-tailed generation. In this work, we fill this gap by analyzing how flow matching models behave under class imbalance and proposing a label-free method to improve their performance in long-tailed regimes.

To overcome these challenges, we propose ***UOT-Reweighted Flow Matching (UOT-RFM)***, a novel framework built on the Unbalanced Optimal Transport (UOT) [6, 25]. Our method leverages mini-batch UOT to construct the conditional vector field and mitigates majority bias through a principled *inverse reweighting* scheme using a *label-free majority score*. This score measures class dominance based on the geometric properties of the data and is defined as the density ratio between the target distribution and the UOT marginal. By incorporating this score into the flow matching objective, UOT-RFM adaptively reweights training samples to emphasize tail modes without requiring class labels.

Our theoretical analysis shows that with first-order correction ($k = 1$), the model recovers the original data distribution, while higher-order corrections ($k > 1$) allow the model to further compensate for majority bias by emphasizing tail samples. Empirically, UOT-RFM significantly outperforms existing flow matching baselines in both tail-class fidelity and the accurate recovery of class proportions, while maintaining competitive performance on standard balanced datasets. Our contributions can be summarized as follows:

- We propose UOT-RFM, the first flow matching framework for long-tailed generative modeling, built on mini-batch Unbalanced Optimal Transport.
- We introduce an unsupervised majority score, derived from the density ratio between the target and UOT marginal distributions, enabling sample-wise reweighting without access to class labels.
- We theoretically and empirically show that higher-order correction using the majority score improves tail sample generation while preserving overall performance.
- To the best of our knowledge, UOT-RFM is the first label-free method for long-tailed generative modeling.

2 Preliminaries

Flow Matching Continuous Normalizing Flows (CNFs) [4, 27] model the dynamics of the probability densities through a *probability density path* $p : [0, 1] \times \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$. The density at time t, denoted as $p_t(\mathbf{x}) := p(t, \mathbf{x})$, transports the initial source distribution (e.g., Gaussian distribution) p_0 to the target data distribution p_1. Specifically, the CNF model is defined by the following Ordinary Differential Equation (ODE), governed by a vector field $\mathbf{v} : [0, 1] \times \mathbb{R}^d \mapsto \mathbb{R}^d$, where $\mathbf{v}_t(\mathbf{x}) := \mathbf{v}(t, \mathbf{x})$:

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \mathbf{v}_t(\mathbf{x}_t), \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ denotes the state variable at time t, and we use the notation $\mathbf{v}_t(\mathbf{x})$ interchangeably with $\mathbf{v}(t, \mathbf{x})$. Then, the associated flow map $\phi_t(\mathbf{x})$ denotes the solution of this ODE with initial condition $\phi_0(\mathbf{x}) = \mathbf{x}$ and the density at time t is given by $p_t = (\phi_t)_{\#}p_0$.

Flow Matching (FM) [27] provides a scalable method for training CNFs. The idea is to train the CNF by minimizing a regression loss $\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta})$ between the parameterized vector field $\mathbf{v}_t^{\theta}$ and the ground-truth vector field $\mathbf{u}_t$ that generates the probability path p_t.

$$\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{x}_t\sim p_t(\mathbf{x}_t)}\|\mathbf{v}_\theta(t, \mathbf{x}_t) - \mathbf{u}_t(\mathbf{x}_t)\|_2^2. \tag{2}$$

However, the marginal ground-truth vector field $\mathbf{u}_t$ is typically intractable. To overcome this, Conditional Flow Matching (CFM) [27, 38] introduces a tractable objective by regressing to a conditional vector field $\mathbf{u}_t(\mathbf{x}_t|\mathbf{z})$. This field generates a conditional probability path $p_t(\mathbf{x}_t|\mathbf{z})$, where $z = (\mathbf{x}_0, \mathbf{x}_1)$

represents sample pairs drawn from a joint distribution (coupling) $\pi(\mathbf{z}) = \pi(\mathbf{x}_0, \mathbf{x}_1)$. The CFM objective is defined as:

$$\mathcal{L}_{\text{CFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{z}\sim\pi(\mathbf{z}),\mathbf{x}_t\sim p_t(\mathbf{x}_t|\mathbf{z})}\|\mathbf{v}_\theta(t,\mathbf{x}_t) - \mathbf{u}_{t|\mathbf{z}}(\mathbf{x}_t|\mathbf{z})\|_2^2. \tag{3}$$

where $t \sim \mathcal{U}[0,1]$. CFM replaces the intractable marginal vector field with a tractable conditional one based on couplings. For example, the conditional probability path $p_t(\mathbf{x}_t|\mathbf{z})$ and the associated conditional vector field $\mathbf{u}_t(\mathbf{x}_t|\mathbf{z})$ can be defined as follows [38]:

$$p_t(\mathbf{x}_t \mid \mathbf{z}) = \mathcal{N}\left(\mathbf{x}_t \mid t\mathbf{x}_1 + (1-t)\mathbf{x}_0 \mid \sigma^2 I\right), \quad \mathbf{u}_t(\mathbf{x}_t \mid \mathbf{z}) = \mathbf{x}_1 - \mathbf{x}_0, \tag{4}$$

where $\sigma > 0$ is a bandwidth hyperparameter. In this case, the marginal probability path and the marginal vector field that generates this path are given by

$$p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t \mid \mathbf{z})\pi(\mathbf{z})d\mathbf{z}, \quad \mathbf{u}_t(\mathbf{x}_t) := \mathbb{E}_{\pi(\mathbf{z})}\left[\frac{\mathbf{u}_t(\mathbf{x} \mid \mathbf{z})p_t(\mathbf{x} \mid \mathbf{z})}{p_t(\mathbf{x})}\right] = \mathbb{E}_{p_t(\mathbf{z}|\mathbf{x}_t)}\left[\mathbf{u}_t(\mathbf{x}_t \mid \mathbf{z})\right]. \tag{5}$$

Initial Coupling in Flow Matching A key component in training flow matching models is the choice of the initial coupling $\pi(\mathbf{z}) = \pi(\mathbf{x}_0, \mathbf{x}_1)$. **The choice of coupling crucially determines the training dynamics of flow matching models**, because the obtained model $\mathbf{v}_t^\theta(\mathbf{x}_t) \approx \mathbf{u}_t(\mathbf{x}_t)$ relies on aggregating the conditional vector field over paired samples $p_t(\mathbf{z}|\mathbf{x}_t)$ (Eq. 5). The original flow matching framework [27] employs an independent coupling between the source and target distributions, i.e., $\pi(\mathbf{x}_0, \mathbf{x}_1) = \mu(\mathbf{x}_0) \otimes \nu(\mathbf{x}_1)$. However, independent couplings often lead to curved trajectories that incur high computational costs during sampling [28]. To improve couplings, recent works adopted the *Optimal Transport (OT)* approaches between mini-batches [31, 38]. The Kantorovich formulation of the Optimal Transport [43] is given by:

$$C_{OT}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)}\left[\int_{\mathcal{X}\times\mathcal{Y}} c(\mathbf{x},\mathbf{y})d\pi(\mathbf{x},\mathbf{y})\right], \tag{6}$$

where $\Pi(\mu,\nu)$ denotes the set of all joint probability measures on $\mathcal{X}\times\mathcal{Y}$ whose marginals are μ and ν respectively. The optimal coupling π is defined as the minimizer of the transport cost $c(\mathbf{x},\mathbf{y})$ between empirical measures of mini-batches from the source samples $\mathbf{x}_0$ and target samples $\mathbf{x}_1$.

3 Method

In this section, we present our model, ***UOT-Reweighted Flow Matching (UOT-RFM)***, that addresses the majority bias of flow matching models on long-tailed distributions. Our model leverages mini-batch UOT coupling, which naturally provides a ***majority score*** for each target sample. Intuitively, we compensate for majority bias by reweighting each target sample with the majority score. In Sec 3.1, we introduce the UOT problem. In Sec 3.2, we introduce our UOT-RFM model.

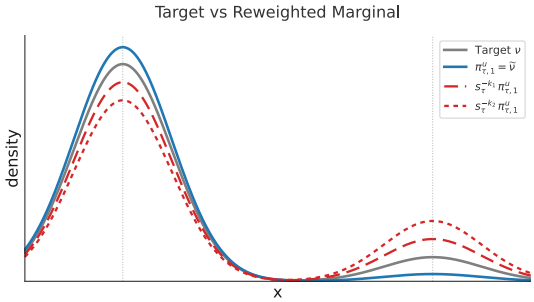
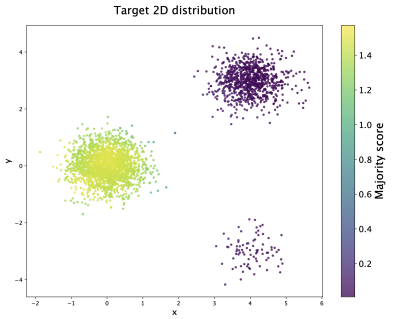


Figure 1: **Comparison of the target data distribution ν and the reweighted marginals $s_\tau^{-k_1}\pi^u_{\tau,1}$ from UOT-RFM** with correction order k (where $1 < k_1 < k_2$). The UOT marginal $\pi^u_{\tau,1}$ down-weights the minority classes. UOT-RFM adaptively upweights minority modes via the majority score s_τ.

3.1 Unbalanced Optimal Transport

Our proposed method builds on the *Unbalanced Optimal Transport (UOT) problem* [6, 25]. In the standard OT (Eq. 6), the marginal distributions of the coupling π are constrained to *exactly* match the source and target distributions, i.e., $\pi_0 = \mu$ and $\pi_1 = \nu$. Although this exact marginal matching is a core principle of OT, it also makes the formulation highly sensitive to outliers [2, 7, 15, 36].

In contrast, the *Unbalanced Optimal Transport* formulation relaxes these hard marginal constraints by introducing divergence penalties between the marginals π_0, π_1 and the source/target measures μ, ν.



τ	Head	Medium-tail	Extreme-tail
1	1.43	0.05	0.04
3	1.31	0.33	0.33
5	1.21	0.53	0.50
10	1.12	0.73	0.72

Figure 2: **Empirical validation of the majority score s_τ. Left:** Visualization on a 3-component 2D Gaussian mixture with imbalanced weights (Head: 0.70, Medium-tail: 0.275, Extreme-tail: 0.025) at $\tau = 1$. The score captures head-tail structure without labels. **Right:** Class-wise average majority scores across varying τ. As τ increases, the spread between head and tail decreases.

This relaxation enables approximate transport, thereby improving robustness to outliers. Formally, the Kantorovich-type UOT formulation is given by:

$$C_{UOT}(\mu,\nu) := \inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})}\left[\int_{\mathcal{X}\times\mathcal{Y}} c(\mathbf{x},\mathbf{y})d\pi(\mathbf{x},\mathbf{y}) + \tau_1 D_{\Psi_1}(\pi_0\|\mu) + \tau_2 D_{\Psi_2}(\pi_1\|\nu)\right], \quad (7)$$

where we assume $c(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ and $\tau_1, \tau_2 > 0$ control the strength of the marginal matching penalties. Here, $\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})$ indicates the set of positive Radon measures on $\mathcal{X}\times\mathcal{Y}$. The terms $D_{\Psi_1}(\pi_0\|\mu)$ and $D_{\Psi_2}(\pi_1\|\nu)$ are two f-divergences that penalize deviations of the coupling marginals π_0, π_1 from the source μ and target ν, respectively. For the convex function Ψ, the f-divergence D_Ψ is formally defined as

$$D_\Psi(\pi_i\|\eta) = \int \Psi\left(\frac{d\pi_i(\mathbf{x})}{d\eta(\mathbf{x})}\right) d\eta(\mathbf{x}) \quad (8)$$

This relaxed formulation allows the optimal UOT coupling π^u (which depends on τ_1 and τ_2) to softly match the marginals, i.e., $\pi_0^u \approx \mu$ and $\pi_1^u \approx \nu$, in contrast to the exact marginal constraints in standard OT. This flexibility allows the coupling to adaptively focus on high-density regions while down-weighting outliers or minority modes that would otherwise incur excessive transport costs.

Moreover, the UOT problem can represent exact matching of one marginal by appropriately setting the divergence penalty. Specifically, if Ψ_i is the convex indicator function ι at $\{1\}$, then $D_\iota(\pi_i\|\eta) = 0$ if $\pi_i = \eta$ a.s., and ∞ otherwise. By setting $\Psi_1 = \iota$ (i.e., $\tau_1 = \infty$), we obtain the **source-fixed UOT problem**, where $\pi_0^u = \mu$ and $\pi_1^u \approx \nu$. In this case, the optimal coupling depends only on the single parameter τ. In our approach, we employ this **source-fixed UOT** formulation to ensure that the initial distribution of the flow matching model aligns exactly with the source distribution.

3.2 Proposed Method

Problem Statement Our goal is to develop a generative model that accurately approximates **long-tailed data distributions, without relying on class labels**. To enable this label-free approach, we assume that *geometrically low-density regions correlate with semantic minority classes*. Our UOT-based approach performs correction based on the geometric property of data x_i, rather than semantic class labels y_i. (This correlation is empirically verified in Fig. 2.) Here, we refer to such geometrically low-density samples as outliers. While a small fraction of outliers may not belong to semantic minorities (e.g., noisy samples) as discussed in Appendix J, we empirically observe that the two largely overlap (Fig. 2 and Appendix E.3).

Formally, we consider a long-tailed dataset $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$, where each $\mathbf{x}_i$ is an input image and $\mathbf{y}^i \in \mathcal{C}$ is its class label. In the long-tailed setting, a small number of classes (**head classes**) dominate with many samples, while most classes (**tail classes**) have only a few, resulting in severe class imbalance. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ denote the set of M classes, ordered by decreasing sample count such that $n_1 \geq n_2 \geq \cdots \geq n_M$, where n_j is the number of training samples in class c_j. The degree of imbalance is quantified by the **imbalance ratio** $\mathcal{I} = n_M/n_1$. Our objective is to faithfully generate all classes, including tails, without using labels $\mathbf{y}^i$.

Majority Score Our method leverages the **mini-batch UOT coupling** π^u and the induced **majority score**, defined as the density ratio s_τ. This score is used to address the majority bias of flow matching models on long-tailed distributions by inversely reweighting each target sample during training. Intuitively, the optimal UOT coupling π^u trades off the transport cost $\int c(\mathbf{x}, \mathbf{y})\, d\pi$ against the marginal penalty $\tau D_{\Psi_2}(\pi_1 \| \nu)$ in Eq. 7. Increasing $D_{\Psi_2}(\pi_1 \| \nu)$ means that the target marginal π_1 deviates further from ν, which relaxes the marginal constraint and allows a lower transport cost. Hence, π^u accepts such a deviation only when the resulting decrease in the transport cost is larger than the increase in the penalty τD_{Ψ_2}. Consequently, π^u tends to concentrate mass on high-density (majority) modes while down-weighting low-density (tail or outlier) modes, which contribute little mass and incur high transport costs (Fig. 1). This inherent geometric mechanism enables UOT to mitigate the influence of outliers effectively [2, 7, 36].

Based on this property, we formally define the ***majority score*** under the source-fixed UOT problem as

$$s_\tau(\mathbf{x}_1) := \frac{d\pi^u_{\tau,1}}{d\nu}(\mathbf{x}_1) > 0, \tag{9}$$

where $\pi^u_{\tau,1}$ denotes the target marginal of the UOT coupling, and $s_\tau : \mathbb{R}^d \to \mathbb{R}_{>0}$ is defined on the target distribution space. Since we employ the source-fixed UOT formulation $\tau_1 = \infty$, the coupling depends only on τ_2. Thus, we simply denote $\tau = \tau_2$ for clarity. Intuitively, the majority score measures how strongly each target sample is emphasized by the UOT coupling. $s_\tau > 1$ indicates emphasized majority samples, while $s_\tau \ll 1$ correspond to down-weighted outlier samples. Here, note that this process is entirely **unsupervised**, i.e., label-free. The down-weighting of outlier modes is conducted by the intrinsic geometric structure of probability distributions (See Appendix A for the formal theoretical relationship between the marginal distributions of π^u_τ and the original source and target distributions).

We demonstrate that the majority score s_τ effectively characterizes the degree of majority on the synthetic dataset. The data distribution is set to a 3-component 2D Gaussian mixture with highly imbalanced weights (Head: 0.70, Medium-tail: 0.275, Extreme-tail: 0.025). The source distribution is a zero-mean Gaussian. (See Appendix E.3 for experiment details.) Fig. 2 shows that the majority score accurately reflects the head-tail structure by yielding higher values for the head class (average of 1.43) and significantly lower values for extreme-tail samples (average of 0.04). Furthermore, it demonstrates how s_τ changes depending on the marginal matching intensity parameter τ. As τ increases, the UOT marginal more closely approximates the data distribution, causing the spread between head and tail scores to decrease as they cluster toward 1. Specifically, the disparity between the head and extreme-tail average scores shifts from 1.43 vs. 0.04 at $\tau = 1$ to a narrower 1.12 vs. 0.72 at $\tau = 10$. This indicates that smaller τ values provide a more discriminative signal for identifying minority samples. The majority score's correlation with semantic head-tail classes extends to real image datasets (Appendix E.3).

UOT-RFM Objective We now introduce our learning objective for modeling long-tailed distributions. Our method consists of two key components: (1) utilizing mini-batch (source-fixed) UOT coupling π^u_τ for pairing samples and (2) rebalancing minority samples through importance correction using majority score $s_\tau(\cdot)$. Our corrected conditional flow matching objective with correction order $k \geq 1$ is defined as follows (Algorithm 1):

$$\mathcal{L}_{\text{ours},k}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{z}\sim\pi^u_\tau(\mathbf{z}),\mathbf{x}_t\sim p_t(\mathbf{x}_t|\mathbf{z})} \left[s_\tau(\mathbf{x}_1)^{-k} \|\mathbf{v}_\theta(t, \mathbf{x}_t) - \mathbf{u}_{t|\mathbf{z}}(\mathbf{x}_t|\mathbf{z})\|_2^2 \right], \tag{10}$$

where $q(t, \mathbf{z}, \mathbf{x}_t)$ represents the joint generative process, with $t \sim \mathcal{U}(0, 1)$, $\mathbf{x}_t \sim p_t(\mathbf{x}_t|\mathbf{z})$, and the conditioning variable $\mathbf{z} = (\mathbf{x}_0, \mathbf{x}_1)$ is sampled from the UOT coupling, i.e., $\mathbf{z} \sim \pi^u_\tau$. The weighting factor $s_\tau(\mathbf{x}_1)^{-k}$ adaptively rebalances majority and minority samples during training. Our method is theoretically grounded by the following bias correction theorem (see Appendix B for proof):

Theorem 3.1. *Let π^u_τ be the optimal source-fixed UOT coupling between μ and ν with $\tau_2 = \tau > 0$ and assume that its target marginal satisfies $\nu \ll \pi^u_{\tau,1}$, i.e., ν is absolutely continuous w.r.t. $\pi^u_{\tau,1}$. Training a flow matching model with π^u_τ yields the biased distribution $p_1 = \pi^u_{\tau,1} \neq \nu$. However, applying the first-order correction (our method with $k = 1$) recovers the true target distribution ν. More generally, UOT-RFM with correction order k generates a distribution $p_1 \propto s_\tau^{-k}\pi^u_{\tau,1} = s_\tau^{-(k-1)}\nu$.*

The assumption $\nu \ll \pi^u_{\tau,1}$ ensures that all target modes, including tail classes, retain nonzero density under the UOT marginal, thereby enabling correction with $s_\tau(\mathbf{x}_1)^{-k}$. Moreover, we impose a **source-**



(a) I-CFM (b) OT-CFM (c) UOT-RFM (Ours)

Figure 3: **Qualitative comparison of generated samples** on CIFAR-10-LT with imbalance ratio $\mathcal{I} = 0.01$. UOT-RFM produces more diverse images compared to other baselines.

Table 1: **Evaluation of marginal distribution matching under the LT→LT setting for the long-tailed benchmarks.** We report FID scores (↓) on CIFAR-10-LT and CIFAR-100-LT with two imbalance ratios: $\mathcal{I} = 0.01$ and $\mathcal{I} = 0.001$.

Model	CIFAR-10-LT		CIFAR-100-LT	
	$\mathcal{I} = 0.01$	$\mathcal{I} = 0.001$	$\mathcal{I} = 0.01$	$\mathcal{I} = 0.001$
I-CFM	14.36 ± 0.20	17.49 ± 0.08	26.06 ± 0.39	32.05 ± 0.21
OT-CFM	17.34 ± 0.23	21.18 ± 0.46	31.20 ± 0.44	38.46 ± 0.28
UOT-CFM	15.20 ± 1.04	18.02 ± 0.23	25.47 ± 0.10	32.36 ± 0.53
UOT-RFM (Ours)	**11.17** ± 0.15	**13.05** ± 0.23	**15.29** ± 0.09	**18.25** ± 0.15

fixed condition on the UOT coupling (i.e., $\pi_0^u = \mu$) to ensure that the initial distribution of the flow matching model remains aligned with the source distribution.

Theorem 3.1 shows that training a flow matching model with UOT coupling (UOT-CFM, [9]) yields a biased generated distribution $p_1 = \pi_{\tau,1}^u \neq \nu$. In particular, the distribution $\pi_{\tau,1}^u$ magnifies the majority modes while suppressing the tail modes. This bias can be corrected by applying inverse weighting with the majority score s_τ. Building on this, our method further addresses the majority bias of standard flow matching models by **over-correction** ($k > 1$) with the majority score. Intuitively, this over-correction amplifies the contribution of geometrically rare samples with $s_\tau < 1$. Note that our goal is to faithfully recover the long-tailed data distribution p_{data}. This over-correction is introduced to compensate for the practical majority bias of standard flow matching induced by optimization error. In contrast to OT-CFM [38], which adopts mini-batch OT coupling, our UOT-based approach provides an **unsupervised estimate of the majority score without requiring additional information such as tail-class label**. Refer to Appendix C for a discussion of related works.

Minibatch UOT Approximation To implement UOT-RFM efficiently, we approximate the UOT coupling π_τ^u using a mini-batch formulation [11]. To be more specific, we adopt the POT library [12] to compute mini-batch UOT with entropic regularization [5, 13]. For each mini-batch of training data $(\{\mathbf{x}_0^i\}_{i=1}^B, \{\mathbf{x}_1^j\}_{j=1}^B)$, the mini-batch coupling $\hat{\pi}_\tau^u$ is computed between empirical measures $\hat{\mu} = \frac{1}{|B|}\sum_i \delta_{\mathbf{x}_0^i}$ and $\hat{\nu} = \frac{1}{|B|}\sum_j \delta_{\mathbf{x}_1^j}$. Then, the score s_τ is estimated as the probability mass ratio:

$$\hat{s}_\tau(\mathbf{x}_1^j) := \frac{\hat{\pi}_{\tau,1}^u}{\hat{\nu}}(\mathbf{x}_1^j) = |B|\hat{\pi}_{\tau,1}^u(\mathbf{x}_1^j). \tag{11}$$

4 Experiments

In this section, we evaluate our model from the following perspectives.

- In Sec 4.1, we evaluate our model on the long-tailed image datasets and analyze the majority bias of flow matching models.
- In Sec 4.2, we assess our model on the standard balanced image datasets, showing that our model remains competitive with small-order correction.

- In Sec 4.3, we conduct ablation studies to investigate the effects of the correction order k and the marginal matching parameter τ.

In each experiment, our model is compared with several flow matching baselines: independent coupling (*I-CFM*, [27, 38]), OT coupling (*OT-CFM*, [31, 38]), and *UOT coupling* (UOT-CFM, [9]). Implementation details are provided in Appendix D.

4.1 Evaluation on Long-Tailed Generation

We evaluate our model across several long-tailed generation benchmarks. First, we conduct a comprehensive evaluation on **CIFAR-10-LT** and **CIFAR-100-LT** [3]. These datasets are constructed by subsampling the balanced CIFAR-10 and CIFAR-100 datasets [21] according to an exponential decaying long-tailed class distribution, i.e., $n_i = \lfloor n_1 \cdot \mathcal{I}^{\frac{i}{M-1}} \rfloor$. The degree of imbalance is quantified by the imbalance ratio $\mathcal{I}$, defined as the ratio between the sample sizes of the most and least frequent classes, i.e., $\mathcal{I} = \min_i\{n_i\}/\max_i\{n_i\}$. Second, to assess the scalability of our model, we test UOT-RFM on large-scale natural image datasets, FFHQ-64 [19] and Places-LT-256 [29].

Long-Tailed Generation We first assess whether our model accurately approximates the true marginal distribution under long-tailed settings. To this end, we consider two evaluation settings, while keeping the training data fixed to the long-tailed dataset:

- **LT→LT**: The test set is also long-tailed (e.g., CIFAR-10-LT test set). This setting assesses how well a model captures the long-tailed distribution.
- **LT→Balanced**: The test set is the original balanced dataset (e.g., CIFAR-10 test set). This setting examines whether a model trained on imbalanced data can recover a balanced distribution. This evaluation setup is often used in supervised LT learning when class labels are available [32, 45].

Our primary evaluation protocol is LT→LT, because our goal is to directly model and approximate the long-tailed data distribution without relying on class supervision. For completeness, we also report LT→Balanced results under an imbalance ratio of $\mathcal{I} = 0.01$.

Table 1 provides FID [17] scores on CIFAR-10 and CIFAR-100 under the **LT→LT setting** for two class imbalance ratios, i.e., $\mathcal{I} \in \{0.01, 0.001\}$. Fig. 3 shows qualitative comparisons of generated samples from baseline flow matching models trained on CIFAR-10-LT with $\mathcal{I} = 0.01$. In both imbalance ratios, UOT-RFM achieves significant improvement in the FID score, demonstrating a more accurate approximation of the long-tailed data distribution. Note that this performance gain comes with minimal computational overhead: UOT-RFM requires only about 7% more training time compared to OT-CFM.

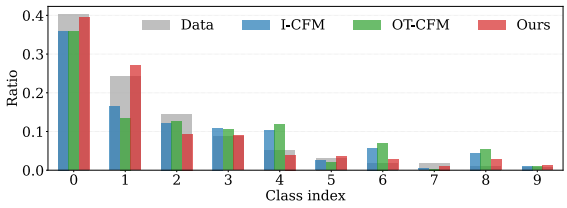
To validate our unsupervised approach, we designed *oracle supervised baselines* that replace our majority score with ground-truth class frequencies: $s_\tau(x_1, y) = p(y)$. We applied this oracle correction to both independent coupling (*O-I-CFM*) and UOT coupling with $\tau = 2$ (*O-UOT-CFM*). On CIFAR-10-LT ($\mathcal{I} = 0.01$), these supervised methods achieve FID scores of 35.35 and 33.03, significantly worse than our unsupervised UOT-RFM (11.17). This demonstrates that naively injecting class-frequency corrections does not satisfy the bias correction property of Theorem 3.1, and highlights the importance of our UOT-based reweighting strategy.

Moreover, Table 2 reports FID scores under the **LT→Balanced** setting with $\mathcal{I} = 0.01$. The trends are consistent with those in the LT→LT evaluation. UOT-RFM (Ours) outperforms all baselines, achieving the lowest FID score—particularly in the more challenging CIFAR-100-LT case. We omit results for the more extreme case of $\mathcal{I} = 0.001$ in this setting. Note that $\mathcal{I} = 0.01$ already represents a highly challenging scenario under a label-free setting. This requires the model to counterbalance a 100:1 disparity between the most and least frequent classes, without access to any class label information.

Table 2: **Evaluation of marginal distribution matching under the LT→Balanced setting for the long-tailed benchmarks.** We report FID scores (↓) with imbalance ratio $\mathcal{I} = 0.01$.

Model	CIFAR-10-LT	CIFAR-100-LT
I-CFM	25.46	24.39
OT-CFM	27.51	29.19
UOT-CFM	24.94	24.05
Ours	**24.06**	**16.83**

Addressing Majority Bias We further investigate how well each model handles majority bias through a detailed class-wise evaluation.



Model	NCRE (↓)	
	CIFAR-10-LT	CIFAR-100-LT
I-CFM	0.84	2.41
OT-CFM	1.02	2.79
UOT-CFM	1.24	3.00
UOT-RFM (Ours)	**0.40**	**1.82**

Figure 4: **Generated class distribution and NCRE evaluation.** The class distribution for CIFAR-10-LT (**Left**) and the quantitative NCRE scores for both datasets (**Right**) demonstrate that our method consistently achieves the best alignment with target long-tailed proportions. These metrics emphasize that UOT-RFM effectively mitigates majority bias.

Table 3: **Quantitative evaluation of mode coverage** under the LT→LT setting with $\mathcal{I} = 0.01$.

Model	CIFAR-10-LT			CIFAR-100-LT		
	Precision (↑)	Recall (↑)	F1 (↑)	Precision (↑)	Recall (↑)	F1 (↑)
I-CFM	0.71	0.29	0.41	0.62	0.29	0.40
OT-CFM	**0.73**	0.24	0.36	**0.73**	0.24	0.36
UOT-CFM	**0.73**	0.29	0.42	**0.73**	0.27	0.40
UOT-RFM (Ours)	0.62	**0.41**	**0.49**	0.69	**0.32**	**0.44**

First, we **compare the class distribution of generated samples to that of the ground-truth test set**. Since our model is *unconditional* (i.e., it does not take class labels as input), we utilize pretrained classifiers on CIFAR-10 and CIFAR-100 to assign proxy labels to generated samples[2]. Fig. 4 shows the generated class distribution on CIFAR-10-LT (see Fig. 7 in Appendix for CIFAR-100-LT). Compared to I-CFM and OT-CFM, our UOT-RFM produces a class distribution that more closely matches the ground-truth long-tailed distribution. To quantify this, we compute the **Normalized Class Ratio Error (NCRE)**, defined as the relative deviation between the generated and true class proportions: $\text{NCRE}_i = \frac{|r_{gen,i} - r_{data,i}|}{r_{data,i}}$, where $r_{gen,i}$ denotes the proportion of generated images assigned to class c_i (via proxy labels) and $r_{data,i}$ is the ground-truth class proportion. Since this metric is normalized by the true proportion, misalignment in tail classes is penalized more heavily. As shown in the table within Fig. 4, UOT-RFM most accurately approximates the target class distribution across both benchmarks. Specifically, for CIFAR-10-LT, the class-average NCRE scores are 0.84 for I-CFM, 1.02 for OT-CFM, 1.24 for UOT-CFM, and 0.40 for UOT-RFM. These results confirm that UOT-RFM significantly reduces distribution error compared to existing flow matching baselines (see Figs 8 and 9 for classwise scores on CIFAR-10-LT and CIFAR-100-LT).

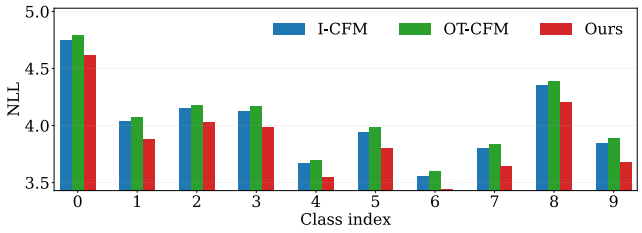
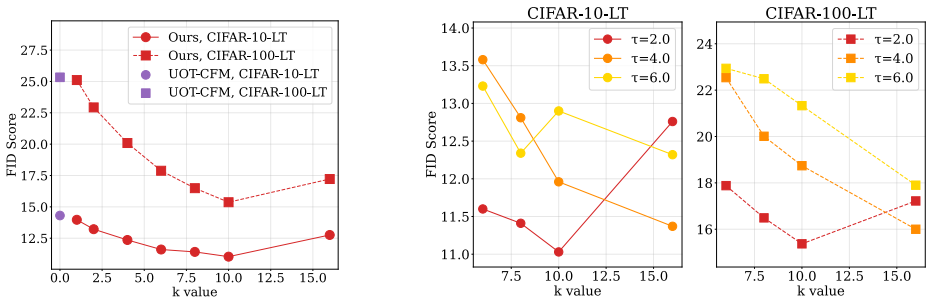


Figure 5: **Class-wise negative log-likelihood of CIFAR-10-LT** data samples on each model. Our method achieves an overall average NLL of 3.88, while I-CFM and OT-CFM achieve 4.02 and 4.06, respectively.

Second, we evaluate the models using Precision, Recall, and F1-score [22], which **provide explicit measurements of mode coverage and balance in sample generation**. Table 3 presents the scores of each flow matching model on CIFAR-10-LT and CIFAR-100-LT with imbalance ratio $\mathcal{I} = 0.01$. Across both datasets, UOT-RFM achieves the best Recall, demonstrating superior coverage of tail modes. While OT-CFM obtains the highest precision metric, UOT-RFM achieves the best F1-score, reflecting a more balanced trade-off between Precision and Recall.

[2] `https://github.com/chenyaofo/pytorch-cifar-models`



(a) Ablation study on k when $\tau = 2.0$ (b) Ablation study on τ across multiple k (UOT-RFM)

Figure 6: **Ablation studies** on the correction order k and the marginal matching strength τ.

Finally, we assess the **fidelity of generated samples from tail classes**. For the quantitative metric, we employ classwise negative log-likelihood (NLL (↓)). Fig. 5 shows the classwise NLL scores on CIFAR-10-LT (see Fig. 12 for CIFAR-100-LT results). UOT-RFM consistently achieves lower NLL across almost all classes, including the tail classes, compared to I-CFM and OT-CFM. For example, the mean NLL on CIFAR-10-LT is 3.88 for UOT-RFM, compared to 4.02 and 4.06 for I-CFM and OT-CFM, respectively (see Fig. 15 for the qualitative comparison of generated images on tail classes).

In summary, these experimental results demonstrate that UOT-RFM more effectively addresses majority bias than existing flow matching models by achieving better alignment with the target class distribution, improved minority class coverage, and higher-fidelity samples on tail modes.

Table 4: **Evaluation of latent flow matching under the LT→LT setting on large-scale LT benchmarks.** We report FID scores (↓) on FFHQ and Places-LT.

Model	FFHQ-64	Places-LT-256
I-CFM	14.37	19.67
OT-CFM	14.24	18.68
UOT-CFM	14.44	**18.08**
Ours	**13.53**	18.12

Latent Space Scalability To evaluate the scalability of UOT-RFM (Ours), we conduct experiments within the Latent Flow Matching framework (LFM) using the FFHQ-64 [19] and Places-LT-256 [29] datasets, under the **LT→LT** setting. Here, FFHQ, which consists of human face images, is considered a naturally imbalanced dataset due to data collection biases: a long-tailed distribution along demographic attributes (e.g., age, ethnicity, accessories) has been reported in the literature on this dataset, even though explicit labels for these attributes are not provided. As shown in Table 4, UOT-RFM consistently outperforms I-CFM and OT-CFM in the latent space, achieving superior FID scores of 13.53 and 18.12, respectively. These results suggest that our label-free reweighting mechanism is sufficiently robust for high-resolution generative tasks (See Appendix I for details).

4.2 Evaluation on Standard Balanced Generation

To further test the general applicability of our method beyond long-tailed distributions, we evaluate UOT-RFM (Ours) on the standard balanced benchmarks: CIFAR-10 and CIFAR-100. Table 5 shows the results. As suggested in Theorem 3.1, UOT-RFM with exact correction ($k = 1$) can be applied to balanced generative modeling. Our model achieves performance comparable to existing flow matching baselines. This result demonstrates that UOT-RFM is not limited to long-tailed settings but also remains competitive on standard balanced distributions.

Table 5: **FID scores (↓) on standard balanced benchmarks.**

Model	CIFAR-10	CIFAR-100
I-CFM	3.78	6.39
OT-CFM	3.64	**6.14**
UOT-CFM	3.62	6.45
Ours	**3.58**	6.54

4.3 Ablation Study

Correction Order k We conduct an ablation study on the correction order k to examine its impact on performance. Fig. 6a shows that introducing correction consistently improves FID scores over UOT-CFM (i.e., UOT-RFM without correction). Interestingly, the best FID scores are achieved at $k = 10 \gg 1$, rather than with the exact correction $k = 1$. We attribute this to the need for stronger upweighting of minority samples in practice, in order to compensate for the exponentially decaying

class distribution under long-tailed settings. Overall, UOT-RFM is robust to correction order k and consistently outperforms other baseline models for all $1 \leq k \leq 16$.

Marginal Matching Intensity τ In addition, we perform an ablation study on the marginal matching intensity $\tau = \tau_2$ in the mini-batch UOT (Eq. 7). The parameter τ controls the degree to penalize the marginal errors in the UOT problem. Hence, increasing τ leads to the more closely matched UOT marginal π_1^u. Consequently, the majority score becomes close to one, $s_\tau(y) \approx 1$, with smaller variance between majority and minority samples (Fig. 2). Fig. 6b shows how the FID scores change according to the different values of τ. When τ is small, even a modest correction order k is sufficient to strongly adjust the generated distribution $p_1 \propto s_\tau^{-(k-1)}\nu$ (Thm. 3.1; see Appendix G.1 for a quantitative analysis). We observe that larger τ values reduce the sensitivity of UOT-RFM to k, stabilizing performance across different settings. However, the best FID scores for long-tailed generation are achieved at a smaller τ, where the model can better adaptively up-weight tail samples.

5 Conclusion

In this paper, we introduced UOT-Reweighted Flow Matching (UOT-RFM) for long-tailed distributions. By leveraging Unbalanced Optimal Transport, we introduced a label-free majority score to correct majority bias through inverse weighting and higher-order corrections. We theoretically justified our reweighting scheme and demonstrated its practical effectiveness across long-tailed benchmarks, where UOT-RFM achieves superior class distribution fidelity, balanced sample generation, and state-of-the-art performance among flow matching models.

Acknowledgements

Funding. H. S. Song and M. J. Gim were supported by National Institute for Mathematical Sciences (NIMS) funded by the Korea government (MSIT) (No. B26810000). J. W. Choi was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00349646, RS-2026-25506577) National Institute for Mathematical Sciences (NIMS) grant funded by the Korea government (MSIT) (No. S26120000), and BK21 FOUR Project (X-AI Driven Immersive Learning Innovation Research and Education Group).

Competing Interests. The authors declare no competing interests.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[4] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[5] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

[6] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274 (11):3090–3123, 2018.

[7] Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[8] Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.

[9] Luca Eyring, Dominik Klein, Théo Uscidda, Giovanni Palla, Niki Kilbertus, Zeynep Akata, and Fabian J Theis. Unbalancedness in neural monge maps improves unpaired domain translation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2UnCj3jeao.

[10] Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2mZSlQscj3. Featured Certification.

[11] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International conference on machine learning*, pages 3186–3197. PMLR, 2021.

[12] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

[13] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

[14] Thomas Gallouët, Roberta Ghezzi, and François-Xavier Vialard. Regularity theory and geometry of unbalanced optimal transport. *arXiv preprint arXiv:2112.11056*, 2021.

[15] Milena Gazdieva, Jaemoo Choi, Alexander Kolesov, Jaewoong Choi, Petr Mokrov, and Alexander Korotin. Robust barycenter estimation using semi-unbalanced neural optimal transport. In *The Thirteenth International Conference on Learning Representations*, 2025.

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[20] Saeed Khorram, Mingqi Jiang, Mohamad Shahbazi, Mohamad H Danesh, and Li Fuxin. 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*, pages 7580–7590, 2024.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

[24] Sangyun Lee, Beomsu Kim, and Jong Chul Ye. Minimizing trajectory curvature of ODE-based generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18957–18973. PMLR, 23–29 Jul 2023.

[25] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[26] Yexiong Lin, Yu Yao, and Tongliang Liu. Beyond optimal transport: Model-aligned coupling for flow matching. *arXiv preprint arXiv:2505.23346*, 2025.

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

[28] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=XVjTT1nw5z`.

[29] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[30] Dogyun Park, Sojin Lee, Sihyeon Kim, Taehoon Lee, Youngjoon Hong, and Hyunwoo J Kim. Constant acceleration flow. *Advances in Neural Information Processing Systems*, 37:90030–90060, 2024.

[31] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28100–28127. PMLR, 23–29 Jul 2023.

[32] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18434–18443, 2023.

[33] Harsh Rangwani, Konda Reddy Mopuri, and R Venkatesh Babu. Class balancing gan with a classifier in the loop. In *Uncertainty in Artificial Intelligence*, pages 1618–1627. PMLR, 2021.

[34] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In *European Conference on Computer Vision*, pages 426–442. Springer, 2022.

[35] Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

[36] Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *arXiv preprint arXiv:2211.08775*, 2022.

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[38] Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[39] Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *European Conference on Computer Vision*, pages 414–430. Springer, 2024.

[40] Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[41] Soobin Um, Beomsu Kim, and Jong Chul Ye. Boost-and-skip: A simple guidance-free diffusion for minority generation. In *Forty-second International Conference on Machine Learning*, 2025.

[42] Adrien Vacher and François-Xavier Vialard. Semi-dual unbalanced quadratic optimal transport: fast statistical rates and convergent algorithm. In *International Conference on Machine Learning*, pages 34734–34758. PMLR, 2023.

[43] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[44] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022.

[45] Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NW2s5XXwXU`.

[46] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Algorithm 1 Minibatch UOT-Reweighted Flow Matching (UOT-RFM)

Input: Empirical or sampleable distributions μ, ν, bandwidth σ, batch size b, initial network $\mathbf{v}_\theta$, sinkhorn target marginal weight τ, weight power scale k.
Initialize: $\tau_1 \leftarrow \infty$ // Source-fixed UOT
while Training **do**
  Sample batches of size b *i.i.d.* from the datasets:
  $\mathbf{x}_0 \sim \mu$; $\mathbf{x}_1 \sim \nu$
  $\pi_\tau^u \leftarrow \mathrm{UOT}(\mathbf{x}_1, \mathbf{x}_0, \tau)$ // Source-fixed UOT $\tau_1 = \infty$, $\tau_2 = \tau$
  $(\mathbf{x}_0, \mathbf{x}_1) \sim \pi_\tau^u$; $t \sim \mathcal{U}(0,1)$
  $\boldsymbol{\mu}_t \leftarrow t\mathbf{x}_1 + (1-t)\mathbf{x}_0$
  $\mathbf{x} \sim \mathcal{N}(\mu_t, \sigma^2 I)$
  Calculate $\hat{s}_\tau(\mathbf{x}_1)$ from Equation (11)
  $\mathcal{L}_{ours}(\theta) \leftarrow \hat{s}_\tau(\mathbf{x}_1)^{-k} \|\mathbf{v}_\theta(t, \mathbf{x}) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2$
  $\theta \leftarrow \mathrm{Update}(\theta, \nabla_\theta \mathcal{L}_{ours}(\theta))$
end while
return $\mathbf{v}_\theta$

A Unbalanced Optimal Transport

Recall that the classical OT problem assumes an exact transport between two distributions μ and ν, i.e., $\pi_0 = \mu, \pi_1 = \nu$. This exact matching constraint makes OT sensitive to outliers [2, 36] and vulnerable to class imbalance [9]. To address these issues, the *Unbalanced Optimal Transport (UOT)* problem [6, 25] relaxes the hard marginal constraints by introducing divergence penalties with regularization parameters $\tau_1, \tau_2 > 0$:

$$C_{UOT}(\mu,\nu) = \inf_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \left[\int_{\mathcal{X}\times\mathcal{Y}} c(\mathbf{x},\mathbf{y})\, d\pi(\mathbf{x},\mathbf{y}) + \tau_1 D_{\Psi_1}(\pi_0 \| \mu) + \tau_2 D_{\Psi_2}(\pi_1 \| \nu) \right], \quad (12)$$

where $\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})$ denotes the set of nonnegative finite measures on $\mathcal{X}\times\mathcal{Y}$. Here, D_{Ψ_1} and D_{Ψ_2} are f-divergences generated by convex functions Ψ_i, penalizing discrepancies between the marginals π_0, π_1 and μ, ν, respectively. Hence, in the UOT problem, the marginals are only *approximately matched* to μ and ν, in the sense that π_0 and π_1 are close to μ and ν with respect to the divergence penalties, rather than being exactly equal as in OT. Intuitively, UOT can be viewed as solving OT between divergence-relaxed marginals π_0, π_1 and the target measures μ, ν [7]. This relaxation provides robustness to outliers [2] and improved adaptability to class imbalance between μ and ν [9].

Similar to the standard OT problem, the UOT problem also admits a *dual formulation* [7, 14, 42]:

$$C_{UOT}(\mu,\nu) = \sup_{u(\mathbf{x})+v(\mathbf{y}) \le c(\mathbf{x},\mathbf{y})} \left[\int_{\mathcal{X}} -\tau_1 \Psi_1^*\Big(-\tfrac{1}{\tau_1} u(\mathbf{x})\Big)\, d\mu(\mathbf{x}) + \int_{\mathcal{Y}} -\tau_2 \Psi_2^*\Big(-\tfrac{1}{\tau_2} v(\mathbf{y})\Big)\, d\nu(\mathbf{y}) \right], \quad (13)$$

where u and v are continuous functions on $\mathcal{X}$ and $\mathcal{Y}$. Here, f^* denotes the *convex conjugate* of f, i.e., $f^*(z) = \sup_{t\in\mathbb{R}}\{tz - f(t)\}$ for $f : \mathbb{R} \to [-\infty, \infty]$.

This dual problem can be further simplified into a *semi-dual* formulation by eliminating u via the optimality condition:

$$C_{UOT}(\mu,\nu) = \sup_{v \in \mathcal{C}(\mathcal{Y})} \left[\int_{\mathcal{X}} -\tau_1 \Psi_1^*\Big(-\tfrac{1}{\tau_1} v^c(\mathbf{x})\Big)\, d\mu(\mathbf{x}) + \int_{\mathcal{Y}} -\tau_2 \Psi_2^*\Big(-\tfrac{1}{\tau_2} v(\mathbf{y})\Big)\, d\nu(\mathbf{y}) \right], \quad (14)$$

where the c-transform of v is defined as

$$v^c(\mathbf{x}) = \inf_{\mathbf{y}\in\mathcal{Y}} \big(c(\mathbf{x},\mathbf{y}) - v(\mathbf{y})\big).$$

Here, v^c corresponds to the optimal potential u given v.

Finally, the relationship between the marginals of the optimal UOT plan π^u and the original source and target distributions can be expressed using the optimal UOT potential v from the semi-dual problem:

Theorem A.1 ([7, 14, 42]). *Let v be a solution of the dual formulation of the UOT problem between the source distribution μ and the target distribution ν. Then, the marginal distributions of the optimal UOT plan π^u satisfy*

$$d\pi_0^u(\mathbf{x}) \;=\; (\Psi_1^*)'\Big(-\tfrac{1}{\tau_1}v^c(\mathbf{x})\Big)\; d\mu(\mathbf{x}) \quad \textit{and} \quad d\pi_1^u(\mathbf{y}) \;=\; (\Psi_2^*)'\Big(-\tfrac{1}{\tau_2}v(\mathbf{y})\Big)\; d\nu(\mathbf{y}). \tag{15}$$

B Proofs of theorem

In this section, we provide the proof of our bias correction theorem (Theorem 3.1) from the main text. Our proof builds on three key lemmas for the standard flow matching model, originally established in [27, 38], which we restate here for completeness.

Lemma B.1 ([38], Theorem 3.1). *The marginal vector field $\mathbf{u}_t$ generates the probability path $p_t(\mathbf{x}_t)$ from initial conditions $p_0(\mathbf{x}_0)$.*

$$p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t \mid \mathbf{z})\pi(\mathbf{z})d\mathbf{z}, \quad \mathbf{u}_t(\mathbf{x}_t) := \mathbb{E}_{\pi(\mathbf{z})}\left[\frac{\mathbf{u}_t(\mathbf{x} \mid \mathbf{z})p_t(\mathbf{x} \mid \mathbf{z})}{p_t(\mathbf{x})}\right] = \mathbb{E}_{p_t(\mathbf{z}|\mathbf{x}_t)}\left[\mathbf{u}_t(\mathbf{x}_t \mid \mathbf{z})\right] \tag{16}$$

Lemma B.2 ([38], Theorem 3.2). *If $p_t(\mathbf{x}_t) > 0$ for all $\mathbf{x}_t \in \mathbb{R}^d$ and $t \in [0, 1]$, then, up to a constant independent of θ, $\mathcal{L}_{\mathrm{CFM}}$ (Eq. 3) and $\mathcal{L}_{\mathrm{FM}}$ (Eq.2) are equal, and hence*

$$\nabla_\theta \mathcal{L}_{\mathrm{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta). \tag{17}$$

Lemma B.3 ([38], Proposition 3.4). *Let the initial sample coupling be $\pi(\mathbf{z}_0, \mathbf{z}_1)$ and define the conditional vector probability path and vector field as in Eq. 4. Then, the corresponding marginal probability path $p_t(\mathbf{x}_t)$ satisfies the boundary conditions $p_0 = \pi_0 * \mathcal{N}(\mathbf{x}|0, \sigma^2 I)$ and $p_1 = \pi_1 * \mathcal{N}(\mathbf{x}|0, \sigma^2 I)$, where $*$ denotes the convolution operator. Furthermore, assuming regularity properties of q_0, q_1, and the optimal transport plan π, as $\sigma^2 \to 0$, the marginal path p_t and field u_t minimize (7), i.e., $\mathbf{u}_t$ solves the dynamic optimal transport problem between π_0 and π_1. Specifically, $p_0 \to \pi_0$ and $p_1 \to \pi_1$ as $\sigma \to 0$.*

Here, we provide a formal statement of Theorem 3.1 and provide its proof.

Theorem B.4 (Theorem 3.1). *Let π_τ^u be the optimal source-fixed UOT coupling between μ and ν with $\tau_2 = \tau > 0$ and assume that its target marginal satisfies $\nu \ll \pi_\tau^u$. Training a flow matching model with π_τ^u yields the biased distribution $p_1 = \pi_1^u \neq \nu$ [9]. However, applying the first-order correction (our method with $k = 1$) recovers the true target distribution ν.*

$$\mathcal{L}_{\mathrm{ours,k=1}}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{z}\sim\pi_\tau^u(\mathbf{z}),\mathbf{x}_t\sim p_t(\mathbf{x}_t|\mathbf{z})}\left[s_\tau(\mathbf{x}_1)^{-1}\|\mathbf{v}_\theta(t,\mathbf{x}_t) - \mathbf{u}_{t|\mathbf{z}}(\mathbf{x}_t|\mathbf{z})\|_2^2\right]. \tag{18}$$

where the majority score $s_\tau(\cdot)$ is defined as $s_\tau(\cdot) := \frac{d\pi_1^u}{d\nu}(\cdot)$. More generally, UOT-RFM with correction order k generates a distribution $p_1 \propto s_\tau^{-k}\pi_{\tau,1}^u = s_\tau^{-(k-1)}\nu$.

Proof. As an overview, the proof relies on two observations: (1) training with π^u yields $p_1 = \pi_1^u$, i.e., the biased UOT marginal (Theorem A.1) and (2) importance reweighting with s_τ^{-1} corrects this bias, since $\nu = s_\tau^{-1}\pi_1^u$ by the Radon–Nikodym derivative. Substituting this correction into the conditional flow matching loss yields Eq. 18, and hence the generated distribution recovers ν.

Formally, Lemma B.3 shows that training a flow matching model with the optimal source-fixed UOT coupling π_τ^u, i.e.,

$$\mathcal{L}_{\mathrm{UOT-CFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{z}\sim\pi_\tau^u,\mathbf{x}_t\sim p_t(\mathbf{x}_t|\mathbf{z})}\|\mathbf{v}_\theta(t,\mathbf{x}_t) - \mathbf{u}_{t|\mathbf{z}}(\mathbf{x}_t|\mathbf{z})\|_2^2. \tag{19}$$

yields a flow matching model whose boundary conditions converge to $p_0 \to \pi_{\tau,0}^u, p_1 \to \pi_{\tau,1}^u$ as $\sigma \to 0$. By Theorem A.1, we have $\pi_{\tau,0}^u = \mu$ and $\pi_{\tau,1}^u \neq \nu$. Therefore, the UOT-CFM model generates a biased distribution.

Moreover, we now show that our UOT-RFM model with the first-order correction recovers the true target distribution. From Theorem A.1, we have $\pi_\tau^u \ll \nu$, so the Radon–Nikodym derivative exists and corresponds to the majority score. By our assumption $\nu \ll \pi_\tau^u$, it follows $\nu = s_\tau^{-1}\pi_1^u$. Therefore,

$$\begin{aligned}
\mathcal{L}_{\mathrm{ours,k=1}}(\boldsymbol{\theta}) &= \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{z}\sim\pi_\tau^u(\mathbf{z}),\mathbf{x}_t\sim p_t(\mathbf{x}_t|\mathbf{z})}\left[s_\tau(\mathbf{x}_1)^{-1}\|\mathbf{v}_\theta(t,\mathbf{x}_t) - \mathbf{u}_{t|\mathbf{z}}(\mathbf{x}_t|\mathbf{z})\|_2^2\right] &(20)\\
&= \mathbb{E}_{t\sim\mathcal{U}[0,1],\mathbf{z}\sim\pi_\tau^u}\left[s_\tau(\mathbf{x}_1)^{-1}\mathbb{E}_{\mathbf{x}_t\sim p_t(\mathbf{x}_t|\mathbf{z})}\left[\|\mathbf{v}_\theta(t,\mathbf{x}_t) - \mathbf{u}_{t|\mathbf{z}}(\mathbf{x}_t|\mathbf{z})\|_2^2\right]\right] &(21)
\end{aligned}$$

Note that the reweighted coupling $s_\tau(\mathbf{x}_1)^{-1}\pi^u_\tau(\mathbf{x}_0, \mathbf{x}_1)$ has the true target distribution ν as its marginal.

$$\int s_\tau(\mathbf{x}_1)^{-1}\pi^u_\tau(\mathbf{x}_0, \mathbf{x}_1)\, d\mathbf{x}_0 = s_\tau(\mathbf{x}_1)^{-1}\pi^u_{\tau,1}(\mathbf{x}_1) = \nu(\mathbf{x}_1). \tag{22}$$

Then, following a similar argument as in the UOT-CFM case, our UOT-RFM model with the first-order correction ($k = 1$) recovers the true target distribution ν. Note that we specifically employ the source-fixed UOT coupling to ensure that the source marginal $\pi^u_{\tau,0} = \mu$ matches exactly with the initial distribution of the flow matching model. More generally, by a similar argument, UOT-RFM with correction order k generates a distribution $p_1 \propto s_\tau^{-k}\pi^u_{\tau,1}$, up to a normalizing constant. □

C Related Works

Generative Models for Long-tailed Data Many real-world datasets follow long-tailed distributions, where a few dominant classes (Head class) contain the majority of samples, while numerous minority classes (Tail class) consist of a smaller number of samples. Generative modeling for long-tailed distributions often fails to learn the tail classes, resulting in low-diversity, low-quality samples. To address this, several GAN-based approaches have been proposed. CB-GAN [33] introduces a regularizer by utilizing a pretrained classifier to balance class learning. gSR-GAN [34] mitigates tail-class mode collapse through group spectral regularization. UTLO [20] encourages head-to-tail knowledge sharing by combining an unconditional low-resolution generator with a conditional high-resolution generator. More recently, diffusion-based methods have been developed for long-tailed generation. CBDM [32] and LTDM [45] improve tail-class quality by transferring knowledge from head to tail classes. In parallel, test-time guidance methods have been introduced. Um et al. [40] uses proxy class labels to guide minority sampling. Um and Ye [39] leverages a self-consistent minority score for minority guidance. Um et al. [41] proposed a guidance-free method for minority sample generation based on variance-boosted initialization and timestep skipping. In contrast, our method is label-free and corrects majority bias directly during training by leveraging the geometric structure of data via Unbalanced Optimal Transport. Table 6 summarizes this distinction. The goals also differ. Inference-time methods steer sampling toward minority regions of an already-trained model to obtain minority samples. UOT-RFM instead trains the model to approximate the entire long-tailed distribution, including both head and tail modes. Furthermore, to the best of our knowledge, our approach is the first flow matching model designed for long-tailed distributions.

Table 6: **Categorization of related methods for long-tailed or minority generation.** Methods are grouped by whether they require class labels and whether the bias is addressed during training or at inference time.

	Training-time	**Inference-time**
Label-based	CB-GAN [33], gSR-GAN [34], UTLO [20], CBDM [32], LTDM [45]	Um et al. [40]‡
Label-free	**UOT-RFM (Ours)**	Um and Ye [39], Um et al. [40]‡, Um et al. [41]

‡Uses proxy labels derived from a minority score rather than ground-truth class labels.

Coupling in Flow Matching A key design choice in flow matching is the coupling between source and target samples. The original framework [27] employs independent coupling. This independent coupling often produces curved trajectories due to flow crossing, increasing numerical errors and sampling cost [24, 30]. To mitigate this, recent works proposed OT-based couplings between mini-batches [31, 38] or trajectory refinement via pretrained models in Rectified Flow [28]. Model-Aligned Coupling [26] dynamically adjusts couplings during training, aligning them with the flow matching model being trained. Eyring et al. [9] introduced the UOT-based couplings to the image-to-image translation, motivated by its adaptability to class imbalance. However, these approaches do not tackle the majority bias in long-tailed generation. Our work complements them by combining UOT-based couplings with a reweighting mechanism that explicitly mitigates majority oversampling. This is the first attempt to utilize the density ratio between the target distribution and the UOT coupling as the

majority score. We employ this score to weight the flow matching objective for long-tailed generation and formally characterize the resulting generated distribution (Theorem 3.1).

D Implementation Details

This section provides the specific implementation details for our experiments on the CIFAR-10/100 and large datasets (FFHQ and Places).

D.1 Experiments on CIFAR

Datasets We use two datasets for our image generation experiments: the standard CIFAR-10/100 dataset and their long-tailed version, CIFAR-10-LT and CIFAR-100-LT. The CIFAR-10-LT/100-LT are generated to simulate class imbalance, following an exponential decay distribution. The number of samples n_i for each class c_i is determined by the formula $n_i = \lfloor n_{\max} \cdot \mathcal{I}^{\frac{i}{M-1}} \rfloor$, where $M \in \{10, 100\}$ is the total number of classes, $n_{\max}$ is the number of samples in the largest class, and the imbalance factor $\mathcal{I}$ is set to 0.01.

Network Architecture We employ the U-Net architecture provided in the `torchcfm` in [38], without any modifications. The architecture uses four resolution levels with two residual blocks per level in both encoder and decoder, linked by skip connections at matching scales. Each block uses 3×3 convolutions with Group Normalization, SiLU activations, and dropout. Down-sampling is performed by stride-2 convolutions, and up-sampling uses nearest-neighbor interpolation followed by a 3×3 convolution.

Training Details All experiments on CIFAR-10/100 follow the default settings of `torchcfm`. We use the `dopri5` ODE solver. For optimization, we use the Adam optimizer with a learning rate of 2×10^{-4}. The model is trained for a total of 400,000 iterations with a batch size of 128. Data preprocessing includes `transforms.RandomHorizontalFlip()` and normalization of pixel values to the range $[-1, 1]$ using `transforms.Normalize(mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5])`. For stable training, we apply a warmup schedule for the first 5,000 iterations, linearly increasing the learning rate from 0 to its target value, and use gradient clipping with an L2-norm threshold of 1.0. For Unbalanced Optimal Transport (UOT), the entropy regularization parameter ϵ is set to 5×10^{-2}, and the source marginal relaxation weight τ_1 is set to infinity.

Method Details The training process of our proposed method is as follows: (1) Sample mini-batches from each distribution. (2) Compute the coupling (transport plan) between the two mini-batches. (3) Determine the weight for each sample based on the computed transport plan. (4) Estimate the vector fields by feeding the coupled sample pairs into the U-Net and compute the weighted loss. (5) Update the network parameters via backpropagation. The specifics of each coupling method are as follows:

- **ICFM:** Uses an independent coupling, assuming the two distributions are independent.
- **OT-CFM:** Computes the transport plan π using the `pot.emd` function and samples pairs according to the normalized probability distribution.
- **UOT-CFM:** Computes the transport plan π^u using the `pot.unbalanced.sinkhorn_knopp_unbalanced` function and samples pairs based on the normalized probabilities.
- **UOT-RFM:** Also uses `pot.unbalanced.sinkhorn_knopp_unbalanced`, but samples only one target for each source sample from the normalized transport plan π^u_τ.

The sample weights are calculated using the column sums of the transport plan π, which corresponds to the empirical measure of the target distribution, denoted as $\tilde{\nu}$. The weight $s_\tau^{-k}(\mathbf{x}_1)$ is defined as Eq. 9. The final loss function is the weighted mean squared error (MSE) between the vector fields: $\mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim\pi^u_\tau}\left[s_\tau^{-k}(\mathbf{x}_1)\|\mathbf{v}_t(\mathbf{x}_0,\mathbf{x}_1) - \mathbf{u}_t(\mathbf{x}_0,\mathbf{x}_1)\|^2\right]$.

Evaluation Metrics To assess the quality of the generated images, we use the Fréchet Inception Distance (FID), Precision, and Recall. FID scores are calculated using the `cleanfid` library. For evaluation against the standard CIFAR-10/100 dataset, we use the library's built-in feature statistics.

Table 7: Total training time (hours) with our computational resources.

Method	CIFAR-10-LT $\mathcal{I} = 0.01$	CIFAR-10-LT $\mathcal{I} = 10^{-3}$	CIFAR-100-LT $\mathcal{I} = 0.01$	CIFAR-100-LT $\mathcal{I} = 10^{-3}$
I-CFM	35.88	35.77	35.63	36.67
OT-CFM	36.56	36.47	36.97	36.74
UOT-CFM	41.20	46.81	48.49	49.27
UOT-RFM	41.34	44.88	48.16	48.71

For CIFAR-10-LT and CIFAR-100-LT, the real data statistics are computed from a long-tailed dataset generated in the same manner as the training set. Precision and Recall are measured based on a widely-used implementation[3], where the real data distribution is also generated identically to the training setup.

Computational Resource and Training Time We provide the computational resource specification and training time table for CIFAR-LT family training. Experiments are conducted on Dell C4140 with dual-socket CPUs advertised as 24 logical processors, about 186.5 GiB of configured DRAM, and four NVIDIA Tesla V100-SXM2-32GB accelerators. Training time (400,000 iterations) is shown in Table 7. Each individual experiment was performed on a single NVIDIA V100 GPU. Although CPU and RAM resources were occasionally shared across parallel runs, the impact on performance and training dynamics was negligible. The computational resources were allocated and managed via a Slurm cluster. We did not require experiments more extensive than those reported in the manuscript for the purpose of hyperparameter searching.

D.2 Experiments on Large-Scale Datasets

Datasets For large-scale image generation experiments, we use FFHQ, and Places365-LT, and ImageNet. FFHQ is used to evaluate generation under natural density variation in face images, while Places365-LT and ImageNet are used to evaluate long-tailed generation under class-level imbalance. Images are resized and normalized to the range $[-1, 1]$, following the preprocessing pipeline used in our implementation.

Network Architecture We use the code base of Latent Flow Matching [8]. Large-scale experiments are conducted in the latent space. We use a pretrained autoencoder as a frozen first-stage model and train the flow model on the encoded latent variables. The generated latent samples are decoded back to image space using the same autoencoder. For the latent vector field, we use the DiT-based architecture implemented in our codebase.

Training Details The model is trained by sampling target latents from the encoded real images and source latents from a standard Gaussian distribution. We use the same latent flow matching training pipeline across FFHQ and Places365-LT. Optimization, checkpointing, and sampling follow the implementation used in Latent Flow Matching [8]; numerical hyperparameters are reported only in the corresponding experimental result tables.

Table 8: Measured time for one training epoch (batch size = 64, latent DiT-L/2) and linear extrapolation to 500 epochs.

Method	**FFHQ-64**		**Places365-LT**	
	1 epoch (s)	Lin. est. 500 epochs (h)	1 epoch (s)	Lin. est. 500 epochs (h)
I-CFM	784.65	≈ 109.0	670.71	≈ 93.2
OT-CFM	788.24	≈ 109.5	673.82	≈ 93.6
UOT-CFM	789.67	≈ 109.7	676.12	≈ 93.9
UOT-RFM	789.81	≈ 109.7	673.26	≈ 93.5

[3] `https://github.com/blandocs/improved-precision-and-recall-metric-pytorch`

Computational Resource and Training Time Experiments ran on a Dell Precision 7875 workstation hosting an AMD Ryzen Threadripper PRO 9965WX CPU (24 cores / 48 threads, 4.2 GHz base and up to 5.4 GHz max boost per AMD), 256 GB of installed system RAM, and an NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU with 96 GB. We provide measured 1-epoch training time and estimated full-epochs(500) time in Table 8. The extrapolation assumes every epoch equals the timed epoch. It excludes full 50K FID sampling that would normally run once at the last epoch. For the purpose of hyperparameter search, we did not require or conduct any experiments beyond those reported in the manuscript.

E Additional Experimental Results on Majority Bias

E.1 Class Distribution

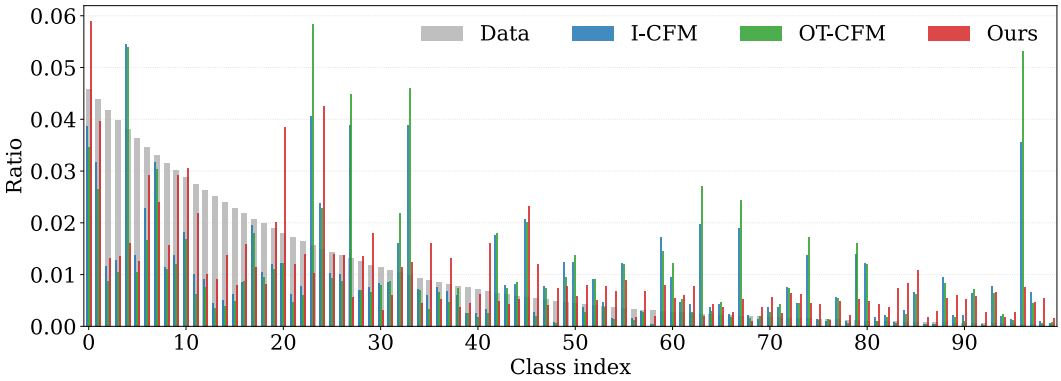


Figure 7: **Generated class distribution on CIFAR-100-LT** with $\mathcal{I} = 0.01$.

Figure 7 visualizes the distribution of the CIFAR-100-LT dataset and the classification results of samples generated by various generative models trained on the CIFAR-100-LT dataset. The classification was performed using a pre-trained classification model (RepVGG-A2) trained on CIFAR100. Ideally, a generative model should produce samples that follow the data's class distribution; however, the results show a notable divergence. Our model, UOT-RFM ($\tau = 2.0, k = 16.0$), can be seen in Figure 7 to better follow the class distribution of the data compared to the baselines I-CFM and OT-CFM.

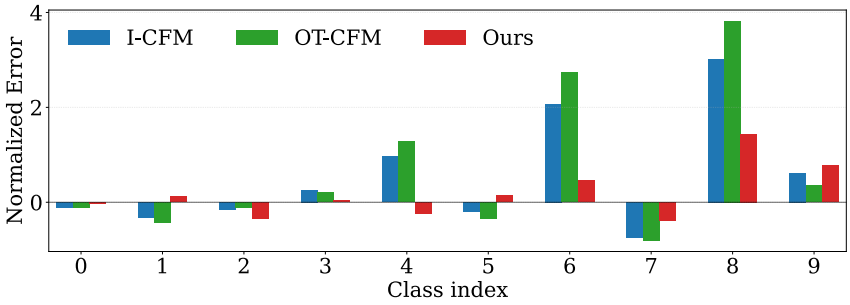


Figure 8: **Signed Normalized Class Ratio Error (Signed NCRE) between generated sample and CIFAR-10-LT. The average NCREs are 0.84 for I-CFM, 1.02 for OT-CFM, and 0.40 for our method.**

In Figure 8, we present a visualization of the Signed Normalized Class Ratio Error (Signed NCRE), calculated as $\frac{r_{gen,i} - r_{data,i}}{r_{data,i}}$ for each class, where a lower absolute value implies better alignment with the data distribution. Our analysis reveals that both I-CFM and OT-CFM exhibit significant discrepancies, while our proposed method shows notably better class distribution alignment (with $\tau = 1.0, k = 10.0$). This is quantitatively supported by the mean of NCRE, which are 0.84, 1.02, and 0.40 for I-CFM, OT-CFM, and our method, respectively.

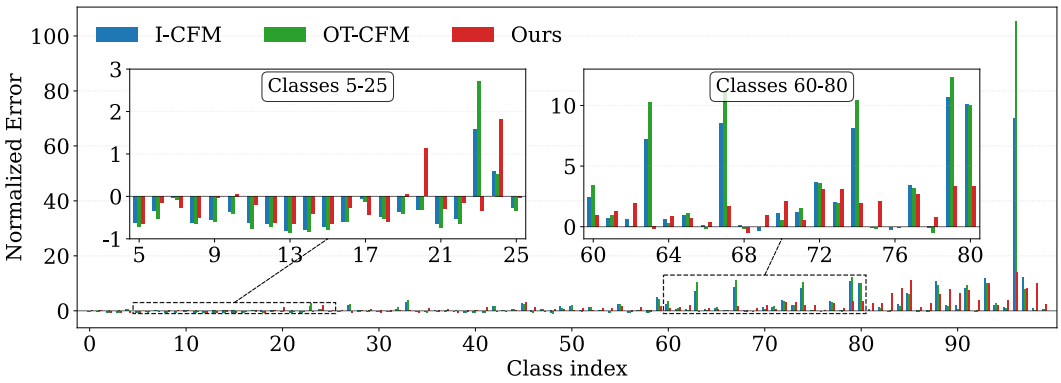


Figure 9: **Signed Normalized Class Ratio Error (Signed NCRE) between generated sample and CIFAR-100-LT**. The average NCREs are 2.41 for I-CFM, 2.79 for OT-CFM, and 1.82 for our method.

We visualized the Signed NCRE of CIFAR-100-LT for each class in Figure 9. While I-CFM and OT-CFM often show large discrepancies, our method ($\tau = 2.0, k = 16.0$) shows a much similar class distribution to that of data. The mean NCRE scores are 2.41 for I-CFM, 2.79 for OT-CFM, and 1.82 for our method, indicating superior distributional fidelity. The two zoomed-in subplots further illustrate the advantage of our approach, particularly in modeling minority classes. In the right subplot (classes 60–80), our method shows substantially lower normalized errors, while I-CFM and OT-CFM display large discrepancies for these rare classes. These visual and quantitative results together confirm that our method more accurately preserves the original class distribution, especially in the challenging tail regions.

E.2 Classwise Negative Log-Likelihood (NLL)

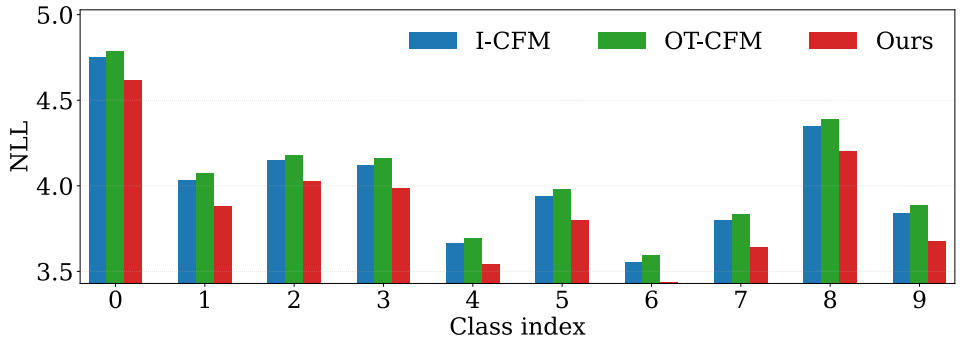


Figure 10: **Class-wise negative log-likelihood of CIFAR-10-LT** data samples on each model. Our method achieves an overall average NLL of 3.88, while I-CFM and OT-CFM achieve 4.02 and 4.06, respectively.

To measure how faithfully each model adheres to the distribution, we also measure the Negative Log-Likelihood (NLL). The log-likelihood of data sample $\mathbf{x}_1$ is computed by solving an ODE at each data sample as shown in the following equation:

$$\log p_1(\mathbf{x}_1) = \log p_0(\mathbf{x}_0) - \int_1^0 \text{div}_t(\mathbf{v}) dt,$$

where integral of divergence is approximated by accumulation of ODE simulation. In our experiments, we use the BPD(bits per dimension) for comparison as NLL $= -\log p_1(\mathbf{x}_1)/(\log(2) \cdot d)$, where d denotes dimension of data.

To further analyze our model's performance on long-tailed distributions, we visualize the class-wise negative log-likelihood (NLL) on the CIFAR-10-LT (imbalance factor $\mathcal{I} = 0.01$) dataset in Figure 5.

Our method achieves the lowest (best) NLL across all ten classes, from index 0 to 9, which indicates that our model can more accurately estimate the distribution within each class. Our method achieves an overall mean NLL of 3.88, while I-CFM and OT-CFM achieve 4.02 and 4.06, respectively.

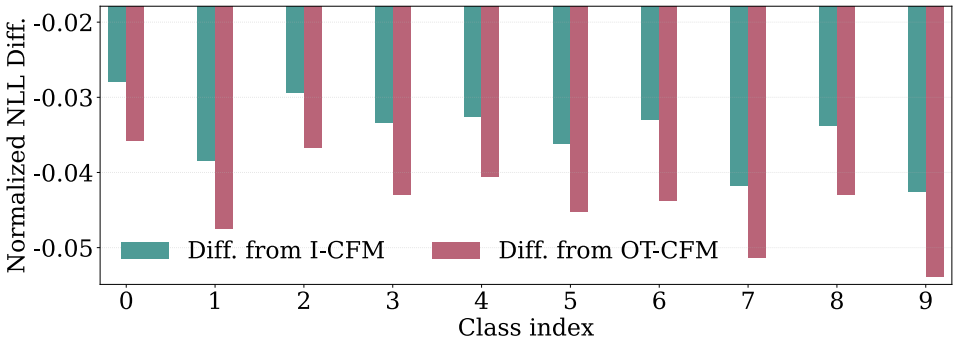


Figure 11: **Normalized difference of class-wise NLL between our model and the baselines on CIFAR-10-LT**.

To provide a more detailed view of this improvement, Figure 11 shows the normalized difference in NLL between our method and the baselines, I-CFM and OT-CFM. The results demonstrate that our NLL values are consistently lower than the baselines, with improvements ranging from a minimum of approximately 2.8% (class 0 and 2, from I-CFM) to a maximum of over 5.0% (class 7 and 9, from OT-CFM). Additionally, the difference is more pronounced in the tail classes. This provides clear evidence that our approach is more effective at capturing the true data distribution, particularly for the underrepresented classes in a long-tailed setting.

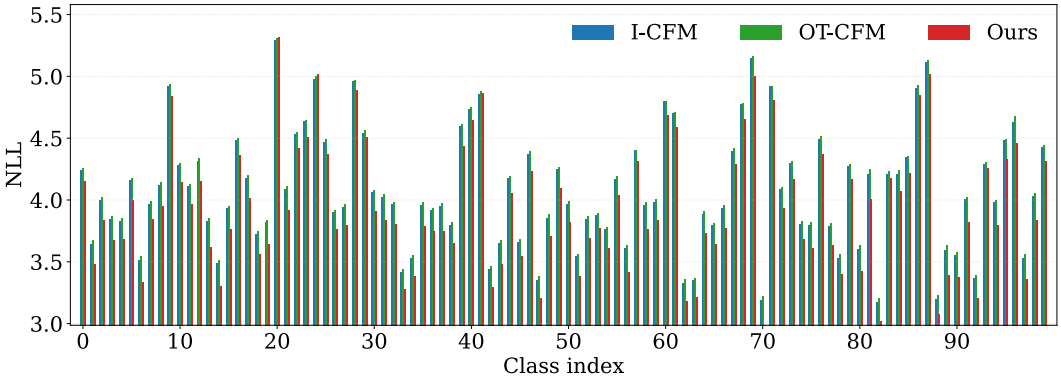


Figure 12: **Class-wise negative log-likelihood of CIFAR-100-LT** data samples on each model. Our method achieves an overall average NLL of 3.94, while I-CFM and OT-CFM achieve 4.08 and 4.10, respectively.

Similarly, we also measured and visualized the class-wise negative log-likelihood for CIFAR-100-LT (imbalance factor $\mathcal{I} = 0.01$). Figure 12 visualizes the average NLL(lower is better) for each class and the mean NLL. Our model achieves 3.94 mean NLL which is lower than I-CFM(4.08) and OT-CFM(4.10) in the CIFAR-100-LT dataset.

We also show the normalized difference between our model and the competing models in Figure 13 to emphasize the gap. Our results show that our method achieves a lower(better) NLL than I-CFM and OT-CFM for most classes.

E.3 Correlation between Majority score and Head-tail Relationship

Synthetic data To evaluate whether the majority-score estimator reflects head–tail relationship, we conducted a synthetic data experiment. Specifically, we constructed a 3-component 2D Gaussian mixture as the target distribution, with clearly imbalanced mixture weights (Head: 0.70, medium-tail: 0.275, extreme tail: 0.025). The component mean vectors are set to $(0, 0)$ for head, $(4, 3)$ for

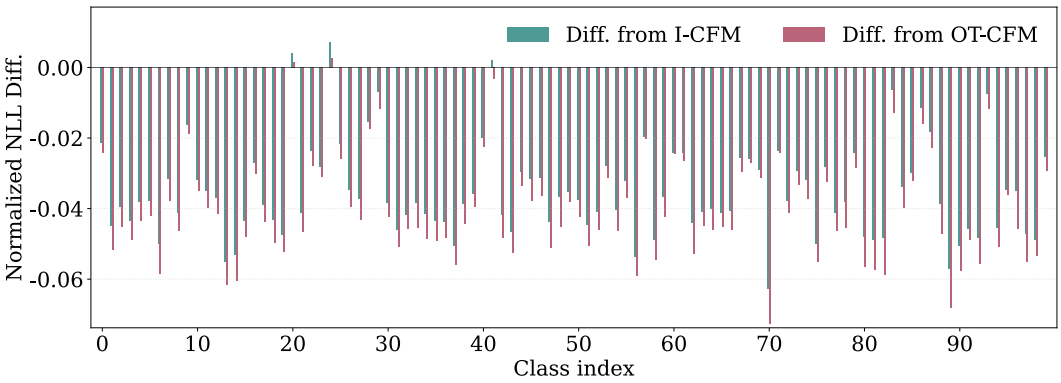


Figure 13: **Normalized difference of class-wise NLL between our model and the baselines on CIFAR-100-LT**.

Table 9: **Weight distribution s_τ^{-k} on Places-LT** in the latent space (1000 samples for head (1 class) and tail classes (91 classes)

Head/Tail	Mean	Min	Max
Head	2.875	0.009	176.7
Tail	3.302	0.008	222.9

medium-tail, and $(4, -3)$ for extreme tail. All components have the same covariance matrix $0.5^2 I$. For the source distribution, we used a single-mode Gaussian with zero mean $(0, 0)$ and covariance $0.5^2 I$. The majority score is computed using the unbalanced Sinkhorn solver [12]. Note that this is identical to how UOT-RFM estimates the majority score between mini-batches.

The results (Fig. 2) show a clear correspondence between the estimated majority score and the true head–tail dominance. Both the qualitative visualization and the average scores exhibit the consistent ordering:

$$\text{head} \;>\; \text{medium-tail} \;>\; \text{extreme-tail},$$

precisely matching the underlying mixture weights. This provides direct evidence that the majority-score estimator correctly captures head–tail structure.

Additionally, the effect of the marginal matching intensity τ is clearly reflected in the results. As τ increases, the optimal marginal π_1 in the UOT objective (Eq. 7) is forced to more closely match the data distribution ν, causing the majority scores s_τ to cluster closer to 1 and reducing the head–tail separation. This trend is confirmed in Table Fig. 2.

Image dataset Table 9 shows weight statistics s_τ^{-k} on Places-LT in latent space. Tail classes receive higher average weights (3.302) than head classes (2.875), confirming that our reweighting correctly up-weights minority samples. The overlapping ranges (Min/Max) indicate that weighting depends on geometric density beyond class labels alone, validating our assumption that geometrically low-density regions are correlated with semantically minority classes, even if they are not identical.

E.4 Sensitivity of the Majority Score to Batch Size

The majority score $\hat{s}_\tau$ is estimated from mini-batch UOT (Eq. 11), so its value may depend on the batch size. We test this sensitivity on CIFAR-10-LT ($\mathcal{I} = 0.01$) with $\tau = 1.0$. We choose $\tau = 1.0$ because it yields a larger spread of the score than the values used in our main experiments ($\tau \in \{2, 4, 6\}$), which makes it a more sensitive setting. We perform two complementary analyses. The first checks whether the overall score distribution shifts with the batch size. The second checks whether each individual sample receives a consistent score across batch sizes.

Distribution-level statistics. We compute the score of every training sample at each batch size and report the summary statistics in Table 10. As the batch size grows from 16 to 2048, the median stays almost unchanged (1.0110 to 1.0103). The standard deviation decreases only slightly (0.0745 to

0.0628), and the maximum decreases slightly while the minimum remains nearly constant. Hence, the overall shape of the score distribution is stable across batch sizes.

Table 10: **Statistics of the majority score across batch sizes** on CIFAR-10-LT ($\mathcal{I} = 0.01$, $\tau = 1.0$).

Batch size	Std	Min	Max	Median
16	0.0745	0.7221	1.1786	1.0110
128	0.0687	0.7183	1.1455	1.0110
512	0.0655	0.7281	1.1341	1.0107
2048	0.0628	0.7294	1.1273	1.0103

Per-sample consistency. A stable distribution does not guarantee that each sample keeps a similar score, since samples could be reshuffled within the same range. To check this directly, let $\hat{s}_\tau^{(m)}(\mathbf{x}_1^i)$ denote the majority score of the i-th training sample estimated by Eq. 11 with batch size $|B| = m$, where $m \in \mathcal{M}$ and $\mathcal{M} = \{16, 32, 64, 128, 256, 512, 1024, 2048\}$. We define the within-sample variance and the across-sample variance as

$$V_{\text{within}} = \frac{1}{N}\sum_{i=1}^{N} \text{Var}_{m\in\mathcal{M}}\left[\hat{s}_\tau^{(m)}(\mathbf{x}_1^i)\right], \qquad V_{\text{across}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \text{Var}_i\left[\hat{s}_\tau^{(m)}(\mathbf{x}_1^i)\right]. \tag{23}$$

V_{within} measures how much the score of a single sample varies due to the batch size. V_{across} measures how much the score varies across samples, which is the head-tail signal that the score is designed to capture. As shown in Table 11, V_{within} is only 5.59% of V_{across}. Therefore, the batch-size-induced fluctuation is small compared to the sample-to-sample variation.

Table 11: **Per-sample consistency of the majority score** across eight batch sizes on CIFAR-10-LT ($\mathcal{I} = 0.01$, $\tau = 1.0$).

Metric	Value
V_{within} (batch-size-induced)	2.221×10^{-6}
V_{across} (sample-induced)	3.975×10^{-5}
Ratio $V_{\text{within}}/V_{\text{across}}$	5.59%

Conclusion. Both analyses indicate that the majority score is robust to the batch size within the range we consider. In particular, a batch size of 128, which we use in all CIFAR experiments, already gives a score distribution close to that of a batch size of 2048.

E.5 Robustness of the Majority Score to Outliers

Numerical stability on extreme outliers. To examine how the majority score behaves on extreme outliers, we construct a controlled dataset by adding pure Gaussian noise samples to CIFAR-10-LT ($\mathcal{I} = 0.01$). The dataset consists of 12,372 normal samples and 6,188 noise samples. We compute the majority score s_τ in both pixel space and latent space with $\tau = 2.0$ and batch size 128. For the latent space, we use the same pretrained autoencoder as in Appendix I.

Table 12: Majority score statistics on CIFAR-10-LT ($\mathcal{I} = 0.01$) with additional Gaussian noise samples.

Space	Group	n	Mean	Std	Min	Max	Median
Pixel	Normal	12372	1.0524	0.0329	0.8992	1.1239	1.0579
	Noise	6188	0.8952	0.0080	0.8661	0.9248	0.8951
Latent	Normal	12372	1.0019	0.0172	0.9039	1.0455	1.0046
	Noise	6188	0.9962	0.0062	0.9658	1.0170	0.9964

Table 12 summarizes the score statistics. In pixel space, the noise samples receive lower scores than the normal samples (mean 0.895 vs. 1.052). However, the scores remain bounded away from zero

(min 0.866), since the entropic regularization in mini-batch UOT prevents the score from collapsing. Thus the inverse weight s_τ^{-k} also remains bounded, and we did not observe training instability. In latent space, the gap becomes much smaller (0.996 vs. 1.002). We attribute this to the pretrained autoencoder, which maps noise inputs to a similar region of the latent space.

Real-world long-tailed dataset. The controlled test above uses synthetic noise as outliers. To examine whether the method remains stable on a realistic long-tailed dataset with natural noise and outliers, we additionally train on ImageNet-LT [29] (256×256) on latent space during 500 epoches. Due to limited computational resources, we compare a single configuration of UOT-RFM ($\tau = 6$, $k = 10$) against I-CFM. As shown in Table 13, UOT-RFM achieves a lower FID than I-CFM. The improvement is modest, and we did not tune τ or k on this dataset. Nevertheless, training remained stable without any additional stabilization technique. This result is consistent with the bounded inverse weights observed in the controlled test. Note that both models are trained as unconditional generators without guidance on a long-tailed dataset with 1,000 classes.

Table 13: FID (↓) on ImageNet-LT.

Model	τ	k	FID
I-CFM	–	–	60.79
UOT-RFM (Ours)	6.0	10.0	**58.30**

E.6 Recall on a Tail Class

Since UOT-RFM assigns large weights to geometrically rare samples, one may ask whether the model generates diverse tail samples or memorizes a few specific tail modes. Recall [23] measures how much of the real data manifold is covered by the generated samples, so it directly reflects the diversity of the generated samples. Tables 19 and 23 report recall over the full dataset. Here, we additionally measure recall on a single tail class to examine diversity within the tail.

We use the dog class of CIFAR-10-LT ($\mathcal{I} = 0.01$), which belongs to the tail and contains 387 training samples. We choose this class because the recall metric relies on k-nearest neighbors of real samples, and classes with too few samples give unreliable estimates. We assign proxy labels to the generated samples with the same pretrained classifier as in Sec. E.1, and compute recall between the generated samples labeled as dog and the real dog samples.

Table 14: **Recall (↑) on the dog class** of CIFAR-10-LT ($\mathcal{I} = 0.01$).

Model	τ	k	Recall
I-CFM	–	–	0.1059
OT-CFM	–	–	0.0698
UOT-CFM	2.0	–	0.1034
UOT-RFM (Ours)	2.0	10.0	**0.1757**

As shown in Table 14, UOT-RFM achieves a recall of 0.1757 on the dog class, which is about 1.7 times higher than the best baseline (0.1059 from I-CFM). This indicates that the generated tail samples cover a larger portion of the real tail distribution, rather than collapsing to a few modes.

F Comparison of ODE Solvers

Table 15: FID comparison between UOT-RFM and OT-CFM across different ODE solvers and NFEs.

Method	**UOT-RFM (↓)**	**OT-CFM (↓))**
dopri5 (NFE: 152/140)	**11.16**	17.58
Euler (5 steps)	38.00	39.13
Euler (10 steps)	19.61	24.99
Euler (20 steps)	14.74	20.07
Euler (50 steps)	12.37	17.84
Euler (100 steps)	11.72	17.37
Euler (200 steps)	11.49	**17.18**
Euler (500 steps)	11.28	17.27

We compared the adaptive ODE solver (*dopri5*) with the Euler solver across a wide range of NFEs (number of function evaluations) (Table 15). Note that the dopri5 solver automatically selects an appropriate number of steps during inference. For example, for randomly selected samples, UOT-RFM and OT-CFM samples are generated under the same hyperparameters atol($1e-5$) and rtol($1e-5$), which yield the comparable NFE of 152 and 140 respectively. With the Euler solver, the FID scores of both models gradually improve as the NFE increases, but UOT-RFM consistently outperforms OT-CFM across all NFEs.

G Additional Ablation Studies

Table 16: **Ablation study on the correction order k and the marginal matching strength τ.** Reported values are FID scores.

	LT→LT				LT→Balanced			
$\tau \backslash k$	1.0	2.0	4.0	8.0	1.0	2.0	4.0	8.0
2.0	13.77	13.41	12.42	11.37	25.02	24.60	24.54	24.76
4.0	14.01	13.78	13.68	12.67	24.94	24.65	24.45	24.37
6.0	14.39	13.72	13.48	12.41	24.91	24.88	24.86	24.90

Table 16 presents an ablation study on the effects of the correction order k and the marginal matching strength τ. All models were trained on the CIFAR-10-LT dataset. The "LT→LT" columns show FID scores measured against the CIFAR-10-LT dataset itself, assessing fidelity to the training distribution. The "LT→Balanced" columns show FID scores using the class-balanced CIFAR10 dataset as a reference, evaluating the generation of a balanced distribution. First, analyzing the LT→LT results, the task is to faithfully replicate the long-tailed training distribution. In this scenario, a clear trend emerges: performance consistently improves as the correction order k increases. For any given value of τ, a larger k leads to a lower (better) FID score. For example, when $\tau = 2.0$, the FID score monotonically decreases from 13.77 at $k = 1.0$ to a superior 11.37 at $k = 8.0$. This indicates that overcorrection ($k > 1$) is consistently beneficial, helping the model to more accurately estimate and represent the target long-tailed marginal distribution. In contrast, the LT→Balanced setting reveals a more complex trade-off. Here, a smaller τ (e.g., 2.0) enables a strong corrective weight but diminishes the sampling probability of minor classes. Conversely, a larger τ (e.g., 6.0) improves the sampling of these classes but flattens the weights, reducing their corrective impact. This necessitates a higher correction order k to induce overcorrection. For instance, with $\tau = 4.0$, increasing k from 1.0 to 8.0 improves the FID score from 24.94 to 24.37. However, excessive overcorrection can overshoot the balanced target, as seen for $\tau = 2.0$, where the FID score worsens from 24.54 ($k = 4.0$) to 24.76 ($k = 8.0$).

Table 17: **FID (↓) scores for CIFAR-10 based datasets**. The **boldface** and underlined values indicate the best and second-best performance. * indicates the best performance of UOT-RFM.

	CIFAR-10			CIFAR-10-LT (0.01)			CIFAR-10-LT (0.001)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	——— 3.78 ———			——— 14.39 ———			——— 17.54 ———		
OT-CFM	——— 3.64 ———			——— 17.49 ———			——— 21.26 ———		
UOT-CFM	3.62	3.72	3.79	14.31	14.37	14.41	19.16	18.13	18.23
ours(1.0)	3.62	3.71	**3.58***	13.97	14.01	14.37	17.15	18.16	17.89
ours(2.0)	3.80	3.60	3.70	13.22	13.76	13.55	16.96	17.12	17.49
ours(4.0)	4.12	3.87	3.82	12.36	13.73	13.62	16.02	16.52	17.34
ours(6.0)	4.51	3.92	3.82	11.60	13.58	13.23	15.13	16.80	16.89
ours(8.0)	4.64	4.63	3.92	11.41	12.81	12.34	13.32	16.24	15.87
ours(10.0)	5.21	4.36	3.92	**11.07***	11.96	12.90	**12.84***	15.05	16.06
ours(16.0)	7.28	4.63	4.20	12.76	11.37	12.32	13.25	13.36	14.64

Table 17 shows the Fréchet Inception Distance (FID) scores on three CIFAR-10 based datasets: CIFAR-10, CIFAR-10-LT ($\mathcal{I} = 0.01$), and CIFAR-10-LT ($\mathcal{I} = 0.001$). Our model consistently achieved the best FID score on each dataset. For the standard CIFAR-10 dataset, our model with $k = 1.0$ and $\tau = 6.0$ achieved the lowest FID score of 3.58, which is an improvement over the previous state-of-the-art model UOT-CFM's best score of 3.62. This improvement can be attributed to the correction of the intra-class feature majority bias. For the CIFAR-10-LT (0.01) and CIFAR-10-LT (0.001) datasets, which are long-tailed distributions, our model achieved the best scores of 11.07 and 12.84, respectively, with $k = 10.0$ and $\tau = 2.0$. These results represent a significant improvement over the baseline models I-CFM and UOT-CFM, which achieved 14.39 and 17.54 respectively.

Table 18: **Precision (↑) scores for CIFAR-10 based datasets.**

	CIFAR-10			CIFAR-10-LT (0.01)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	0.46			0.68		
OT-CFM	0.47			**0.71**		
UOT-CFM	0.46	0.46	0.47	0.67	0.67	0.67
ours(1.0)	0.46	0.46	0.47	0.65	0.65	0.65
ours(2.0)	0.47	0.46	0.47	0.64	0.65	0.66
ours(4.0)	0.47	0.46	0.46	0.63	0.65	0.65
ours(6.0)	0.47	0.46	0.46	0.65	0.65	0.65
ours(8.0)	0.47	0.47	0.46	0.63	0.64	0.64
ours(10.0)	0.47	0.47	0.46	0.61	0.62	0.62
ours(16.0)	**0.48**	0.47	0.46	0.63	0.62	0.63

Table 19: **Recall (↑) scores for CIFAR-10 based datasets.**

	CIFAR-10			CIFAR-10-LT (0.01)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	0.38			0.29		
OT-CFM	0.38			0.24		
UOT-CFM	0.38	0.38	0.38	0.28	0.29	0.28
ours(1.0)	0.38	0.38	0.38	0.27	0.28	0.28
ours(2.0)	0.38	0.39	0.38	0.29	0.28	0.28
ours(4.0)	0.38	0.38	**0.39**	0.29	0.29	0.28
ours(6.0)	0.38	0.38	**0.39**	0.28	0.28	0.28
ours(8.0)	0.38	0.38	0.38	0.33	0.30	0.29
ours(10.0)	0.38	0.38	0.38	0.38	**0.41**	**0.41**
ours(16.0)	0.36	0.38	**0.39**	0.32	0.32	0.31

Table 20: **F1 (↑) scores for CIFAR-10 based datasets.**

	CIFAR-10			CIFAR-10-LT (0.01)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	0.42			0.41		
OT-CFM	0.42			0.36		
UOT-CFM	0.42	0.42	0.41	0.40	0.42	0.40
ours(1.0)	0.42	0.42	0.42	0.38	0.39	0.39
ours(2.0)	0.42	0.41	0.41	0.40	0.40	0.40
ours(4.0)	0.42	0.42	**0.43**	0.40	0.41	0.40
ours(6.0)	0.42	0.41	0.42	0.40	0.40	0.40
ours(8.0)	0.42	0.42	0.41	0.46	0.41	0.40
ours(10.0)	0.42	0.42	0.41	0.47	**0.49**	**0.49**
ours(16.0)	0.41	0.42	0.42	0.43	0.43	0.42

In addition to the FID scores, we conducted a quantitative analysis using precision and recall metrics, with the results summarized in Tables 18 and 19. The precision scores generally remain comparable (CIFAR-10) to or lower (CIFAR-10-LT) than the baselines, while the recall scores show marked improvements (CIFAR-10-LT) in some cases. This behavior is particularly evident on the CIFAR-10-LT (0.01) dataset. For example, our model with $k = 10.0$ and $\tau = 4.0$ achieves a recall of 0.41, a significant improvement over the best baseline score of 0.29 from I-CFM and UOT-CFM. This enhancement in recall highlights our model's increased capability to generate diverse samples that cover the full spectrum of the data distribution, especially the minority classes.

The F1 score, which harmonizes precision and recall, serves as a comprehensive metric for evaluating generative models, and our model consistently shows better performance in cases where recall is improved, as shown in Table 20. Specifically, on the CIFAR-10-LT (0.01) dataset, our models such as a setting with $k = 10.0$ and $\tau = 4.0$ achieve an F1 score of 0.49, which is higher than the best baseline score of 0.42 from UOT-CFM. This demonstrates that even if there is no gain in precision, the increase in recall from our method's ability to better capture minority features leads to a more balanced and representative learned distribution overall.

Table 21 presents the FID scores for CIFAR-100-based datasets. Our model demonstrates a significant performance boost on the long-tailed datasets compared to the baselines. On CIFAR-100-LT ($\mathcal{I} = 0.01$), our best score is 15.38 (with $k = 10.0, \tau = 2.0$), a substantial improvement over the best baseline of 25.33 from UOT-CFM. For the even more imbalanced CIFAR-100-LT ($\mathcal{I} = 0.001$) dataset, our model achieves a score of 18.40 (with $k = 16.0, \tau = 2.0$), which is a large improvement over the best baseline score of 31.83. This demonstrates our method's ability to effectively correct the majority bias present in these imbalanced datasets.

However, on the balanced CIFAR-100 dataset, our model's FID score is slightly higher than that of the competitive models. The best FID score for our method is 6.54, slightly underperforming the best

Table 21: **FID (↓) scores for CIFAR-100 based datasets.**

	CIFAR-100			CIFAR-100-LT (0.01)			CIFAR-100-LT (0.001)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	———	6.39	———	———	25.56	———	———	31.86	———
OT-CFM	———	**6.14**	———	———	31.90	———	———	38.37	———
UOT-CFM	6.48	6.63	6.45	26.96	25.78	25.33	33.32	31.89	31.83
ours(1.0)	6.77	6.54*	6.55	25.11	25.84	25.07	32.67	32.29	31.98
ours(2.0)	6.79	6.65	6.79	22.93	24.68	24.62	31.23	31.73	31.13
ours(4.0)	7.27	6.69	6.67	20.09	23.45	23.14	28.70	30.62	30.37
ours(6.0)	7.68	6.93	7.06	17.88	22.54	22.94	26.01	29.13	30.21
ours(8.0)	8.70	7.20	7.29	16.49	20.01	22.49	23.04	28.43	28.91
ours(10.0)	10.12	7.60	7.18	**15.38***	18.74	21.33	20.04	25.35	28.18
ours(16.0)	15.90	8.73	7.57	17.22	16.00	17.90	**18.40***	22.26	26.19

baseline score of 6.14 from OT-CFM. This suggests that for a balanced dataset, a large correction order k may not be optimal. Instead, very delicate hyperparameter tuning is likely required to achieve superior results. It is also worth noting that our approach, based on UOT-CFM, exhibits a relatively high FID score on CIFAR-100, where UOT-CFM itself has a score of 6.45. This can be attributed to the influence of UOT coupling. We leave the further refinement of the UOT coupling's effect on balanced datasets as a direction for future work.

Table 22: **Precision (↑) scores for CIFAR-100 based datasets.**

	CIFAR-100			CIFAR-100-LT (0.01)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	———	0.41	———	———	0.62	———
OT-CFM	———	**0.43**	———	———	**0.73**	———
UOT-CFM	0.41	0.41	0.41	0.72	**0.73**	0.72
ours(1.0)	0.41	0.41	0.42	0.72	**0.73**	0.72
ours(2.0)	0.41	0.41	0.41	0.71	0.72	0.72
ours(4.0)	0.42	0.42	0.41	0.71	0.72	0.71
ours(6.0)	0.42	0.42	0.41	0.69	0.71	0.72
ours(8.0)	0.41	0.41	0.41	0.69	0.71	0.72
ours(10.0)	0.42	0.41	0.41	0.68	0.71	0.71
ours(16.0)	**0.43**	0.42	0.41	0.65	0.69	0.70

Table 23: **Recall (↑) scores for CIFAR-100 based datasets.**

	CIFAR-100			CIFAR-100-LT (0.01)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	———	**0.37**	———	———	0.29	———
OT-CFM	———	0.36	———	———	0.24	———
UOT-CFM	**0.37**	0.36	0.36	0.28	0.27	0.27
ours(1.0)	0.36	0.36	0.36	0.27	0.27	0.27
ours(2.0)	0.36	0.36	0.36	0.28	0.28	0.29
ours(4.0)	0.35	0.36	0.35	0.27	0.28	0.29
ours(6.0)	0.35	0.35	0.36	0.28	0.29	0.29
ours(8.0)	0.34	0.36	0.36	0.29	0.30	0.28
ours(10.0)	0.33	0.35	0.36	0.30	0.30	0.29
ours(16.0)	0.30	0.34	0.36	0.30	**0.32**	0.31

Table 24: **F1 (↑) scores for CIFAR-100 based datasets.**

	CIFAR-100			CIFAR-100-LT (0.01)		
model(k) \ τ	2.0	4.0	6.0	2.0	4.0	6.0
I-CFM	———	**0.39**	———	———	0.40	———
OT-CFM	———	**0.39**	———	———	0.36	———
UOT-CFM	**0.39**	0.38	0.38	0.40	0.40	0.40
ours(1.0)	0.38	0.38	**0.39**	0.40	0.40	0.40
ours(2.0)	0.38	0.38	0.38	0.40	0.40	0.41
ours(4.0)	0.38	**0.39**	0.38	0.40	0.41	0.41
ours(6.0)	0.38	0.38	**0.39**	0.40	0.41	0.41
ours(8.0)	0.37	0.38	0.38	0.41	0.42	0.40
ours(10.0)	0.37	0.38	0.38	0.41	0.42	0.40
ours(16.0)	0.36	0.38	0.38	0.41	**0.44**	0.42

Tables 22, 23, and 24 present the precision, recall, and F1 scores for the CIFAR-100-based datasets. Our method shows a clear improvement in recall, especially on the CIFAR-100-LT ($\mathcal{I} = 0.01$) dataset. In the best case, with $k = 16.0$ and $\tau = 4.0$, our model achieves a recall score of 0.32, which is notably higher than the best baseline recall of 0.29 from I-CFM. This recall improvement indicates our model's enhanced ability to generate more diverse samples, particularly for the minority classes.

The F1 score, which balances precision and recall, also demonstrates our method's superiority in this setting. On the CIFAR-100-LT ($\mathcal{I} = 0.01$) dataset, our model with $k = 16.0$ and $\tau = 4.0$ achieves an F1 score of 0.44. This is a clear improvement over the best baseline F1 score of 0.40 from I-CFM and UOT-CFM, showing that our method's emphasis on minority features leads to a more balanced and accurate generative process.

However, our method shows slightly lower results on the CIFAR-100. In this case, our approach requires more delicate coupling and tuning rather than aggressive re-weighting to perform optimally, similar to the CIFAR-10 case.

G.1 Relationship between the correction order and marginal intensity

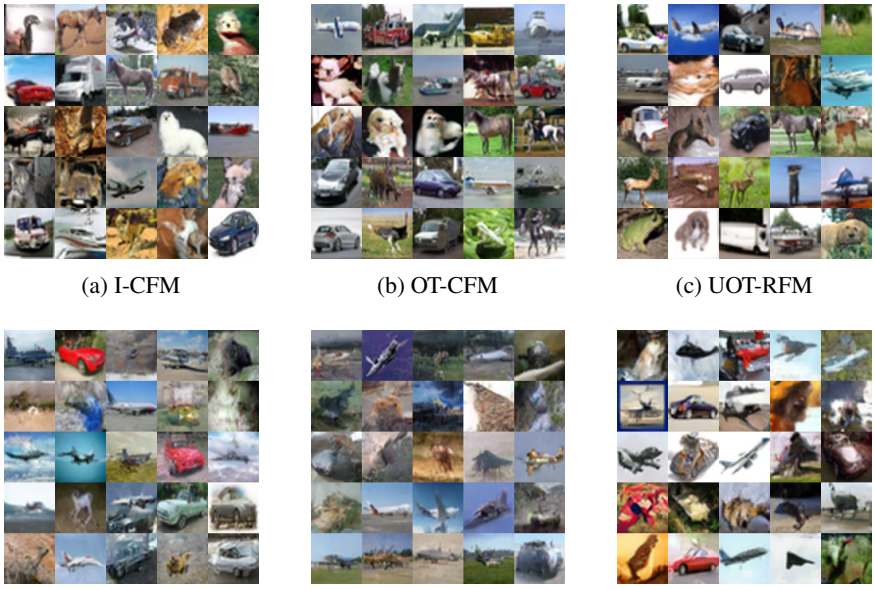
In our ablation studies, the best FID scores are often achieved at a large correction order such as $k = 10$ or $k = 16$ (Tables 17 and 21). Here, we show that the large absolute value of the optimal k is partly explained by the marginal matching intensity τ. As shown in Eq. 7, a larger τ penalizes the deviation of the target marginal $\pi^u_{\tau,1}$ from ν more strongly. Thus, the majority score s_τ concentrates near 1, and its standard deviation becomes small. In this case, the weight s_τ^{-k} also has a small spread unless k is large.

To verify this, we compute the standard deviation of the majority score and of the weight s_τ^{-k} on CIFAR-10-LT ($\mathcal{I} = 0.01$) for various τ and k with batch size 128. Table 25 shows the results. As τ increases from 0.1 to 2.0, the score standard deviation decreases from 0.4006 to 0.0360. Accordingly, a larger k is required to reach a weight with a similar spread. For example, the weight standard deviation reaches about 1 at $k = 4$ for $\tau = 0.5$, while it requires $k = 16$ for $\tau = 2.0$. Since we use $\tau \in \{2, 4, 6\}$ in our many experiments, the optimal k appears large in absolute terms. Therefore, k and τ should be interpreted jointly through the spread of the resulting weight, rather than separately.

Table 25: **Standard deviation of the majority score s_τ and the weight s_τ^{-k}** for various τ and k on CIFAR-10-LT ($\mathcal{I} = 0.01$).

τ	Score std	Weight std			
		$k = 2$	$k = 4$	$k = 10$	$k = 16$
0.1	0.4006	5.5603	252.15	1.13×10^8	8.25×10^{13}
0.3	0.1919	0.6887	3.0922	368.75	77231.5
0.5	0.1270	0.3529	1.0405	17.827	405.04
1.0	0.0689	0.1609	0.3727	1.8262	7.8681
2.0	0.0360	0.0776	0.1650	0.5245	1.1693

H Additional Qualitative Examples



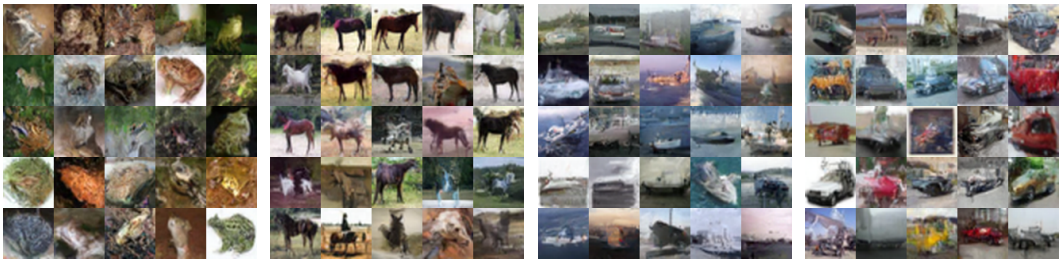
(a) I-CFM (b) OT-CFM (c) UOT-RFM

(d) I-CFM (e) OT-CFM (f) UOT-RFM

Figure 14: **CIFAR image generation results.** The first row shows images randomly generated from models trained on the balanced CIFAR10 dataset. The second row shows images from models trained on the CIFAR-10-LT dataset.

To qualitatively verify our models, we include images actually generated by each model in this paper. Figure 14 shows 25 completely randomly sampled generated images, without cherry-picking. The first row displays generated images from each model trained on the balanced CIFAR-10 dataset, and the second row shows generated images from each model trained on the long-tailed CIFAR-10-LT (0.01) dataset. Our method, UOT-RFM, utilizes hyperparameters $\tau = 1.0$ and $k = 1.0$ when trained on CIFAR-10, and $\tau = 1.0$ and $k = 10.0$ when trained on CIFAR-10-LT.

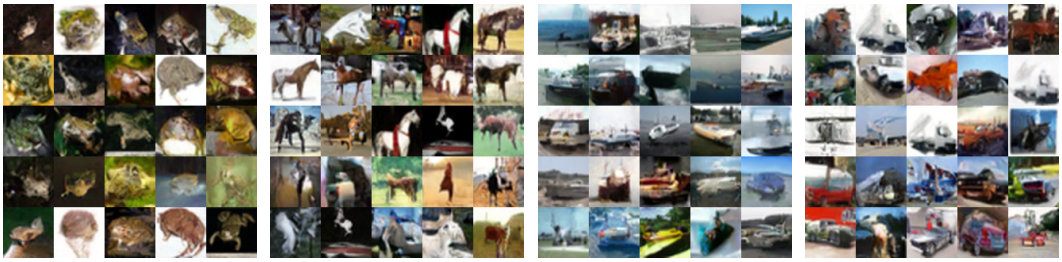
On the CIFAR-10 dataset, our model appears comparable to the baseline models I-CFM and OT-CFM. However, on the CIFAR-10-LT dataset, I-CFM and OT-CFM tend to generate images that are somewhat blurry and noisy. In contrast, our generated images are relatively clean and distinct.



(a) I-CFM, frog (b) I-CFM, horse (c) I-CFM, ship (d) I-CFM, truck



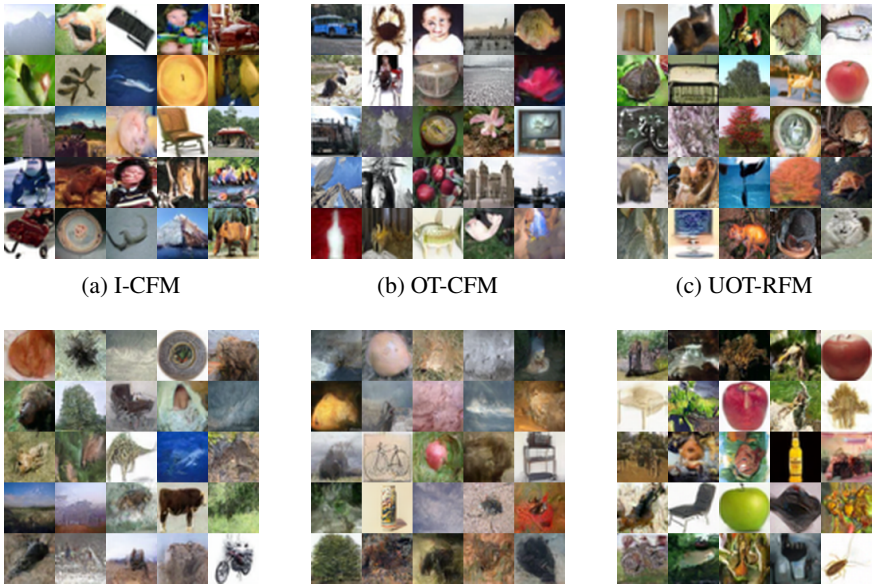
(e) OT-CFM, frog (f) OT-CFM, horse (g) OT-CFM, ship (h) OT-CFM, truck



(i) UOT-RFM, frog (j) UOT-RFM, horse (k) UOT-RFM, ship (l) UOT-RFM, truck

Figure 15: **Tail-classes CIFAR image generation results.** The classified images from models trained on the CIFAR-10-LT dataset.

Figure 15 visualizes the generated images for the tail classes of the CIFAR-10-LT dataset. Specifically, the first column shows images generated for class 06 (frog), the second column for class 07 (horse), the third column for class 08 (ship), and the last column for the last tail class (truck). The samples in each grid image are selected by the top-confident values from a classification model, which was pre-trained on the balanced CIFAR-10 dataset. Observing the results, the images generated by I-CFM and OT-CFM show a tendency to be relatively noisy. In contrast, our method, UOT-RFM, yields images that are cleaner and exhibit greater diversity within the class.

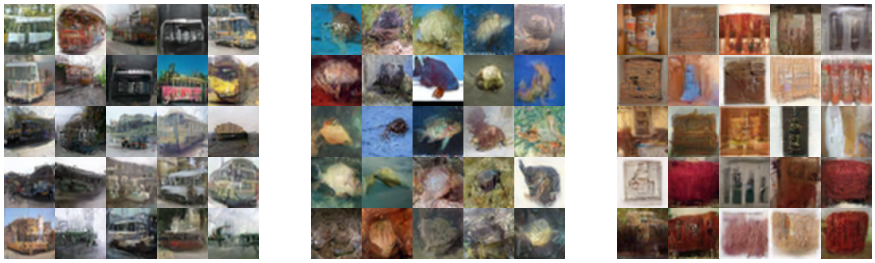


(a) I-CFM (b) OT-CFM (c) UOT-RFM

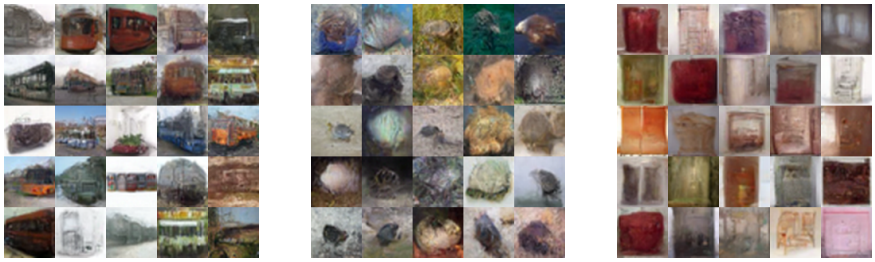
(d) I-CFM (e) OT-CFM (f) UOT-RFM

Figure 16: **CIFAR image generation results.** The first row shows images randomly generated from models trained on the balanced CIFAR100 dataset. The second row shows images from models trained on the CIFAR-100-LT dataset.

Figure 16 shows 25 completely randomly sampled images. The first row displays images generated when trained on the balanced CIFAR-100 dataset, and the second row shows images generated when trained on the long-tailed CIFAR-100-LT (0.01) dataset. The overall trend observed in CIFAR-100 is similar to that in CIFAR-10. Specifically, on the balanced CIFAR-100 dataset, our UOT-RFM method generates images that appear comparable to those from I-CFM and OT-CFM. However, when trained on the long-tailed CIFAR-100-LT, the generated images from I-CFM and OT-CFM tend to be blurry, whereas the images produced by UOT-RFM are noticeably cleaner and more distinct.



(a) I-CFM, streetcar (b) I-CFM, turtle (c) I-CFM, wardrobe



(d) OT-CFM, streetcar (e) OT-CFM, turtle (f) OT-CFM, wardrobe



(g) UOT-RFM, streetcar (h) UOT-RFM, turtle (i) UOT-RFM, wardrobe

Figure 17: **Tail-classes CIFAR image generation results.** The classified images from models trained on the CIFAR-100-LT dataset.

Figure 17 visualizes the generated images for three specific tail classes from the CIFAR-100-LT dataset as examples. The first column shows images generated for class 81 (streetcar), the second column for class 93 (turtle), and the last column for class 94 (wardrobe). The samples in each grid image are selected by the top-confident values of a classification model. In this comparison, I-CFM and OT-CFM, particularly the latter, show a pronounced tendency to produce blurry and noisy images. In contrast, our UOT-RFM method produces images that are noticeably sharper and capture a better variety of features, such as shape, within each class.

I Latent Flow Matching

Table 26: **Evaluation of latent flow matching under the LT→LT setting on large-scale long-tailed benchmarks.** We report FID scores (↓) on FFHQ and Places-LT.

Model	FFHQ-64	Places-LT-256
I-CFM	14.37	19.67
OT-CFM	14.24	18.68
UOT-CFM	14.44	**18.08**
UOT-RFM (Ours)	**13.53**	18.12

To demonstrate the scalability of our method to latent models and its generalization capability across high-resolution, naturally imbalanced, and diverse datasets, we conducted experiments on FFHQ[19] and Places-LT[29] within the Latent Flow Matching framework[8]. FFHQ, which consists of human face images, is considered a naturally imbalanced dataset due to human genetic variations and data collection biases. We use a 64×64 resolution for FFHQ. In contrast, Places-LT are artificially imbalanced datasets constructed by sampling from Places[46] dataset, following a Pareto distribution (PDF: $\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$) with $\alpha = 6$ and $x_m = 5$. Consequently, the number of samples per class ranges from 5 to 4,980 for Places-LT. These datasets are evaluated at a 256×256 resolution. Additionally, our method (UOT-RFM) uses the following hyperparameters: $\tau = 4, k = 4$ for FFHQ-64; $\tau = 2, k = 10$ for Places-LT-256. In this setting, the models perform training and inference in the latent space encoded by a pre-trained VAE encoder, while the remaining algorithmic configurations remain consistent with those used in the pixel-space experiments. As summarized in Table 26, our UOT-RFM achieves the best FID score of 13.53 on the FFHQ-64 dataset, outperforming both I-CFM (14.37), OT-CFM (14.24), and UOT-CFM (14.44). Furthermore, UOT-RFM outperforms I-CFM and OT-CFM on the Places-LT-256 dataset, reaching an FID of 18.12 compared to 19.67 and 18.68, respectively, and performs comparably to UOT-CFM (18.08). These findings suggest that our proposed model is sufficiently effective for generative tasks within the latent space across diverse data distributions.

J Limitations

Training-time Method. While this study focuses on training-time reweighting, UOT-RFM remains inherently orthogonal to inference-time guidance mechanisms. Exploring the synergy with inference-time guidance, particularly by adapting our UOT-based logic for test-time control, remains a natural extension of this work.

Hyperparameter dependency. Our method requires tuning the hyperparameters k and τ. However, the tuning cost and failure risk are minimal as UOT-RFM consistently outperforms all baselines across a wide range of values, as demonstrated in our ablation studies (Fig. 6). Furthermore, k and τ are tightly coupled through the spread of the majority score s_τ, meaning a full 2D grid search is unnecessary. Since the optimal values are governed by the imbalance intensity of dataset, such as the imbalance ratio $\mathcal{I}$, a robust starting point can be selected based on the our experimental results.

Sample Complexity. Our majority score is estimated by *mini-batch UOT*, so the estimate $\hat{s}_\tau$ depends on the batch size. In general, the mini-batch UOT plan approximates the full UOT plan more accurately with larger batch sizes. Our analysis in Appendix E.4 shows that the estimated score is stable across batch sizes from 16 to 2048, both in its overall distribution and for individual samples. However, this analysis is conducted on CIFAR-10-LT only.

Geometric outliers and semantic minorities. Our majority score assumes that geometrically low-density samples correspond to semantic minorities. This assumption may not hold for a small fraction of samples, such as noisy or out-of-distribution samples. Our stress test shows that the inverse weights of such samples remain bounded (Appendix E.5). We did not observe training instability caused by these samples in our experiments. Using a feature extractor with richer semantic information may further reduce this issue.

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 have fully reflected the claims made in the Abstract and Introduction through the methodology, theoretical proofs, and empirical evaluations provided within the manuscript. Specifically, these contributions are supported by Algorithm 1, Theorem 3.1, Equations 9-10, Figures 7, 4, 8, 9, 5, 10, 11, 12, 2, 6, and Tables 1, 2, 3, 4, 15, 16, 17, 20, 21, 24, 25, 26, and 5.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We represent the scope of our method's objectives within the main text of the manuscript and provide a comprehensive discussion of our work's limitations in Appendix J (Limitations section).

   Guidelines:

   - The answer [N/A] 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**

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

   Answer: [Yes]

   Justification: The assumptions and complete proof for Theorem 3.1 (Bias Correction Theorem) are provided in Section 3.2 and Appendix B, respectively. The formal derivation relies on established Conditional Flow Matching properties detailed in Lemmas B.1, B.2, and B.3.

   Guidelines:

   - The answer [N/A] 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: Comprehensive implementation details are provided in Appendix D, where all hyperparameters for the experiments are appropriately specified. Other experimental configurations follow the cited frameworks, specifically Tong et al. (CFM) [38] and Latent Flow Matching [8], ensuring that our work is easily reproducible.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: All source code for the proposed method and baselines is provided in the supplemental material, utilizing publicly available datasets: CIFAR-10/100, FFHQ, ImageNet, and Places. We have included detailed instructions for accessing the raw data as well as the scripts required to generate the long-tailed (LT) variants from the original datasets, as described in Appendix D.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Comprehensive experimental settings, including network architecture, optimization details, and UOT-specific parameters (ϵ, $\tau_1 = \infty$), are provided in Appendix D. Hyperparameters for each dataset (CIFAR-10/100-LT, FFHQ-64, Places-LT-256), specifically k and τ, are detailed in Section 4.3 and Tables 16, 17, and 21.

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: We report the mean and standard deviation for our primary experimental results in Table 1, using the $\pm$ notation to demonstrate statistical significance. While we acknowledge that it was not feasible to conduct multiple runs for every configuration in our extensive ablation studies due to limited computing resources, the statistical results provided in our main benchmarks (Table 1) would be sufficient to confirm the stability and reliability of the proposed method.

Guidelines:

- The answer [N/A] 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: We provide specific details regarding the computational resources and execution time for our experiments in Appendix D. We further clarify that the full research project did not require experiments more extensive than those reported in the manuscript, even for the purpose of hyperparameter searching.

   Guidelines:

   - The answer [N/A] 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: Our research conforms to the NeurIPS Code of Ethics by addressing and mitigating majority bias in generative models, thereby contributing to the development of more responsible machine learning systems. Furthermore, this study utilizes only established,

publicly available datasets—specifically CIFAR-10, CIFAR-100, FFHQ, ImageNet, and Places-LT—and does not involve experiments with human subjects or the collection of sensitive personal information, ensuring no violations of privacy or ethical standards.

Guidelines:

- The answer [N/A] 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: By mitigating majority bias and enhancing the representation of underrepresented classes, our framework promotes fairness in generative modeling. However, as with all high-fidelity generative models, there is a potential risk that the improved generation quality could be misused to create synthetic or fake images for disinformation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our research focuses on foundational algorithmic advancements using standard public benchmarks and does not involve the release of high-risk, large-scale models or sensitive datasets that necessitate specialized safeguards. While we acknowledge the general risks inherent in generative AI, our work remains strictly within an academic scope and does not introduce novel safety threats beyond those present in existing generative frameworks.

Guidelines:

- The answer [N/A] 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**

    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: Our codebase for the main CIFAR experiments is based on the work of Tong et al. [38] (`https://github.com/atong01/conditional-flow-matching?tab=MIT-1-ov-file`) under the MIT License, while the implementation for auxiliary latent experiments follows Dao et al. [8] (`https://github.com/VinAIResearch/LFM?tab=AGPL-3.0-1-ov-file`) and utilizes the AGPL-3.0 License accordingly. In addition, we have provided citations for the sources of all datasets in the main text.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: We provide the sources for all original datasets and explicitly describe the methodology (Cao et al. [3] and Liu et al. [29]) for constructing the long-tailed (LT) variants in Section 4.1 and Appendix D. The scripts for data downloading, LT dataset generation, and model implementation are provided as anonymized supplemental material, which includes an appropriate license to ensure accessibility and reproducibility.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This research does not involve crowdsourcing experiments or research with human subjects. All experiments were conducted using publicly available benchmark datasets, and the study adheres to the NeurIPS Code of Ethics and established research integrity standards.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This research does not involve crowdsourcing or any form of research with human subjects. All experiments were performed using standard, publicly available benchmark datasets. therefore, institutional review board (IRB) approval was not required for this study.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: None of the core methods proposed in this research or implemented in the provided source code involve Large Language Models (LLMs) as an original or non-standard component. LLMs were not utilized in any capacity that impacts the core methodology, scientific rigor, or originality of the work.

    Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.