FLOW MATCHING FOR SOFT-LABEL DEBIASING IN LONG-TAILED DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation compresses large-scale datasets into compact synthetic data to reduce storage and training costs. In long-tailed scenarios, soft labels generated during distillation suffer from entangled bias originating from the imbalanced distillation model and synthetic images. Existing calibration methods typically assume this bias as a linear additive component and apply static statistical shifts to the logits. However, the distribution shift induced by class imbalance is highly nonlinear, rendering linear adjustments insufficient for disentangling semantic relationships from frequency-induced bias. This paper proposes Flow Matching Label Transport (FMLT), a framework that models soft-label debiasing as a continuous transport process on probability manifolds. We formulate the transition from biased soft-label distributions to unbiased target distributions using ordinary differential equations. By learning a time-dependent vector field conditioned on synthetic image features, FMLT transports the biased labels along nonlinear trajectories without relying on additive assumptions. Experiments on CIFAR-10/100-LT and ImageNet-1k-LT demonstrate that FMLT improves performance across various distillation baselines and data budget settings. In extreme long-tailed scenarios, the nonlinear transport mechanism restores inter-class semantic topologies and improves tail-class accuracy.

1 INTRODUCTION

Dataset distillation aims to synthesize a compact surrogate dataset that preserves the essential knowledge of a large-scale original dataset, thereby reducing the computational and storage overhead associated with model training Wang et al. (2018); Zhao et al. (2021a); Cazenavette et al. (2022). Recent methodological advancements have significantly improved the information density of synthetic data Yin & Shen (2024); Shao et al. (2024a); Xiao & He (2024). Within these frameworks, soft labels have emerged as a critical component, preserving inter-class semantic relationships and providing richer supervision signals compared to hard one-hot labels Hinton et al. (2015); Sucholutsky & Schonlau (2021); Xiao & He (2024). However, existing distillation methods predominantly operate under the assumption of class-balanced distributions Zhao et al. (2021a); Cazenavette et al. (2022). Real-world data universally exhibits long-tailed distributions, where head classes contain abundant samples while tail classes are sparsely represented Cui et al. (2019); Kang et al. (2020). Directly applying standard distillation techniques to long-tailed datasets introduces systematic bias, leading to degraded performance on tail classes and optimization instability Jiang et al. (2025); Wu et al. (2026).

Recent studies have investigated the entangled nature of soft-label bias in long-tailed dataset distillation Zhao et al. (2025a); Jiang et al. (2025); Cui et al. (2026). In particular, long-tailed distillation can inherit bias from both the learning process and the synthetic data themselves: experts trained on imbalanced data provide biased gradients and unreliable supervision for tail classes, while the distilled images may further induce distributional distortion in the resulting soft labels Zhao et al. (2025a); Jiang et al. (2025). Jiang et al. Jiang et al. (2025) further identify two distinct sources of soft-label bias, originating from the distillation model and the distilled images, respectively, and formulate their correction through an adaptive soft-label alignment mechanism. More generally, long-tailed recognition methods often compensate for class-frequency-induced bias through post-hoc logit or margin calibration, shifting predictive scores according to class priors or class-dependent corrections Wang et al. (2023); Chen & Su (2023). While such strategies have demonstrated substantial

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

empirical improvements, an additive or globally parameterized correction can be restrictive when the distortion is strongly feature dependent. Long-tailed training not only induces systematic overconfidence toward head classes Chen & Su (2023), but can also deteriorate the representation and supervision available to tail classes Zhao et al. (2025a); Cui et al. (2026). Consequently, the resulting soft-label deformation is unlikely to be fully characterized by a uniform displacement in logit space: confidence bias, semantic ambiguity, and tail-class representation degradation may interact nonlinearly. A simple linear shift may therefore fail to preserve fine-grained inter-class relations, potentially leading to suboptimal decision boundaries and distorted semantic topology in the calibrated soft-label space.

To address this limitation from a geometric perspective, we model a soft label as a probability vector residing on the categorical probability simplex Δ^{C-1} and regard the collection of soft labels as a structured probability space. Under this view, calibrating a biased soft-label distribution toward an unbiased target distribution can be naturally interpreted as a probability transport problem. This perspective is closely related to continuous probability-flow models, where a time-dependent vector field progressively transports one distribution into another Lipman et al. (2023). Recent developments in Riemannian Flow Matching further extend this idea to non-Euclidean spaces by explicitly accounting for the underlying geometry when constructing probability paths and vector fields Chen & Lipman (2024), while categorical Flow Matching has demonstrated that probability distributions supported on the simplex can be modeled through geometry-aware continuous transformations Williams et al. (2026). In contrast, conventional logit adjustment applies a largely global correction determined by class-frequency statistics Wang et al. (2023), without explicitly modeling the local geometry of individual soft-label configurations. This distinction becomes particularly important when head and tail classes exhibit overlapping visual semantics: a global shift may be unable to distinguish genuine semantic similarity from frequency-induced confidence bias Chen & Su (2023). We therefore argue that bias rectification should instead be formulated as a continuous, nonlinear, and feature-conditioned transport process. Such a formulation allows the transport field to adapt to the local geometry of the soft-label space, progressively disentangling frequency-induced bias from semantic relations while preserving the intrinsic relational structure among classes.

Based on this geometric perspective, we propose Flow Matching Label Transport (FMLT), a novel soft-label debiasing framework operating within a continuous ordinary differential equation (ODE) framework. Instead of relying on additive decomposition, we formulate the debiasing process as a continuous transport process on the probability manifold. We define a time-dependent vector field that dictates the transport path from the biased soft-label distribution to the unbiased target distribution. By utilizing synthetic image features as conditional inputs, FMLT learns to regress this vector field, constructing nonlinear trajectories that smoothly transport each biased soft label to its unbiased counterpart. This approach preserves the inter-class semantic topology, as the transport trajectory is sensitive to the specific feature information of the input image. By integrating the vector field using standard ODE solvers, FMLT produces calibrated soft labels that are independent of the long-tailed prior. The proposed method functions as a standalone module, requiring no modifications to the underlying distillation architecture or optimization objectives.

The main contributions of this paper are summarized as follows: 1) We identify the theoretical limitations of the linear additive bias assumption in existing long-tailed dataset distillation methods. By modeling soft labels as distributions on a probability manifold, we provide a geometric interpretation of the nonlinear deformation caused by long-tailed distributions. 2) We propose Flow Matching Label Transport, a continuous ODE-based soft-label debiasing framework. By learning a time-dependent vector field conditioned on synthetic image features, FMLT transports biased label distributions along nonlinear trajectories, disentangling semantic information from frequency-induced bias. 3) Evaluations on CIFAR-10/100-LT and ImageNet-1k-LT demonstrate that FMLT improves performance across multiple distillation baselines and data budget settings. Improvements in extreme long-tailed scenarios validate the effectiveness of nonlinear transport in restoring inter-class semantic topologies.

2 Related Work

Dataset distillation condenses large datasets into compact synthetic counterparts for efficient model training Wang et al. (2018). Early methods based on gradient or trajectory matching require costly

optimization, limiting scalability Zhao et al. (2021b). Distribution matching reduces this cost by aligning feature distributions between synthetic and real data Zhao & Bilen (2023). Recent frameworks, including SRe^{2}L, G-VBSM, and EDC, further enable distillation at the ImageNet scale Yin et al. (2023b); Shao et al. (2024b;c). However, these methods mainly target balanced datasets, and their direct application to long-tailed data can propagate imbalance into synthetic images and soft labels Zhao et al. (2025b).

Real-world classification data often exhibit long-tailed distributions, commonly addressed by reweighting, resampling, or logit adjustment Menon et al. (2021). LTDD first systematically studied dataset distillation under long-tailed distributions and revealed biased distillation supervision for tail classes Zhao et al. (2025b). Recent work further attributes soft-label bias to both imbalanced distillation models and distribution shifts in synthetic images Jiang et al. (2025). Existing calibration applies post-hoc additive adjustment in logit space Jiang et al. (2025). However, such static corrections cannot fully capture feature-dependent nonlinear bias, especially when head and tail classes share similar visual semantics.

Flow Matching learns continuous probability transport by regressing time-dependent vector fields along predefined probability paths Lipman et al. (2023). Its extension to non-Euclidean geometries further enables geometry-aware transport on structured spaces. In contrast, existing long-tailed soft-label calibration remains essentially single-step and static. Viewing soft labels as points on the probability simplex enables debiasing to be formulated as continuous transport. By conditioning the vector field on synthetic-image features, the resulting trajectory can adapt to local semantic structure, providing a nonlinear alternative to additive logit calibration.

3 METHOD

3.1 NONLINEAR ANALYSIS OF SOFT-LABEL DEFORMATION

Existing methodologies in long-tailed dataset distillation approximate the observed soft-label bias through a linear additive decomposition. Specifically, it is assumed that the observed posterior probability $p_{obs}(y|x)$ can be represented as the summation of the target unbiased posterior $p_{target}(y|x)$ and distinct bias terms $\epsilon_T(y|x)$ and $\epsilon_I(y|x)$ originating from the teacher model and the synthetic images, respectively. Based on this linear assumption, post-hoc calibration strategies apply a global scalar shift to the logits to balance class-wise confidence. However, this linear additive assumption is theoretically insufficient for modeling the underlying data generation process under extreme class imbalance. The distortion of soft labels induced by long-tailed distributions is highly nonlinear. The overconfidence in head classes and the feature collapse in tail classes are intertwined, creating a complex deformation that cannot be reversed by a simple linear translation. When head and tail classes share overlapping visual features, linear shifts fail to distinguish genuine semantic overlap from frequency-induced confidence bias, often resulting in suboptimal class boundaries and disrupted semantic topologies within the soft labels.

To fundamentally address this limitation, we model soft labels as distributions residing on a probability manifold. In this geometric context, the observed biased soft-label distribution and the desired unbiased target distribution occupy distinct regions on the manifold. The objective of bias calibration is equivalent to transporting probability mass from the biased region to the unbiased region. Existing linear calibration methods correspond to rigid, global translations on this manifold, which ignore the local curvature and topological structure. To preserve the inter-class semantic relationships encoded in the soft labels, a continuous and nonlinear transport mechanism that adapts to the local geometric structure is required. This observation motivates the formulation of soft-label debiasing as a continuous transport process on the probability manifold.

3.2 CONTINUOUS FORMULATION OF FLOW MATCHING LABEL TRANSPORT

The Adaptive Soft-label Alignment (ADSA) method Jiang et al. (2026), representing existing linear adjustment techniques, applies a global scalar shift to the logits. As illustrated in Figure 1, this static operation corresponds to a rigid translation on the probability manifold. It fails to reshape the local geometric structure and cannot disentangle genuine semantic overlap from frequency-induced bias. In contrast, the proposed FMLT formulates the debiasing process as a continuous transport mechanism. The vector field, conditioned on image features, adapts to the local curvature of the

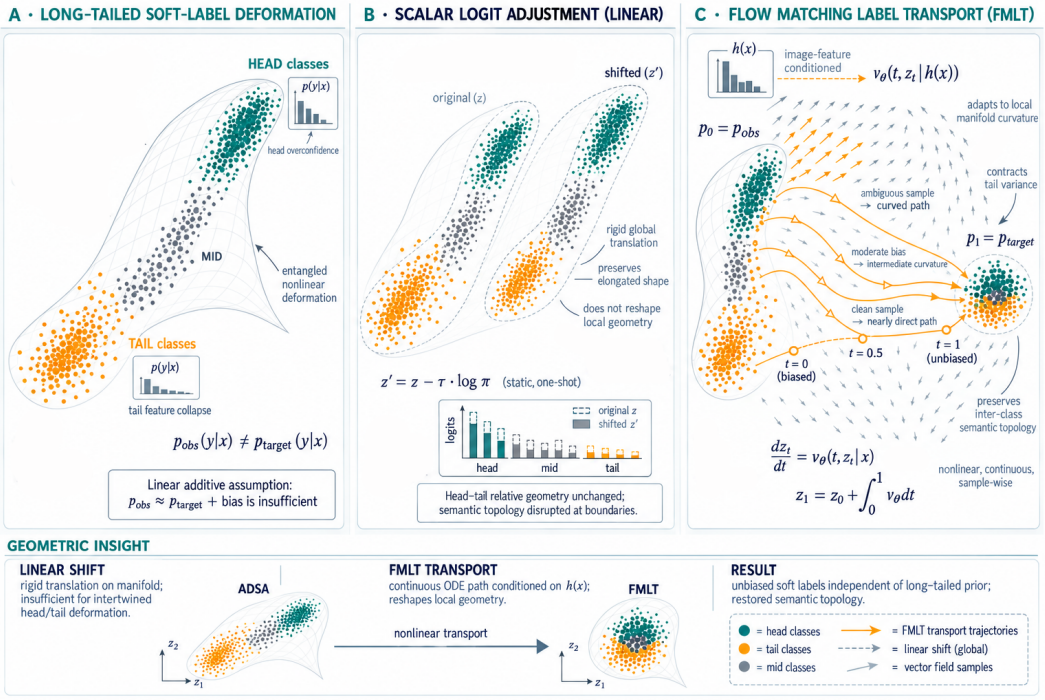

Figure 1: Geometric comparison of soft-label calibration methods. (A) Long-tailed distributions induce nonlinear deformation, causing head overconfidence and tail feature collapse. (B) Scalar logit adjustment applies a rigid global translation, which is insufficient for intertwined head and tail deformations and disrupts semantic topologies. (C) The proposed FMLT utilizes a continuous ODE path conditioned on image features to transport probability mass, adapting to local manifold curvature and restoring inter-class semantic topology.

manifold. This nonlinear trajectory contracts the variance of tail classes and preserves the inter-class semantic topology, transporting the biased distribution to the target unbiased distribution without relying on additive assumptions.

We propose Flow Matching Label Transport, a framework that models the debiasing process as a continuous ordinary differential equation (ODE) Hale (2009). Instead of relying on an additive decomposition, we define a time-dependent probability path $p_t(y|x)$ for $t \in [0, 1]$, where $p_0(y|x)$ corresponds to the observed biased soft-label distribution $p_{obs}(y|x)$, and $p_1(y|x)$ corresponds to the target unbiased distribution $p_{target}(y|x)$. The transition from p_0 to p_1 is governed by a time-dependent vector field $v_t(y|x)$.

Formally, the evolution of the soft-label distribution is described by the continuity equation. Let $z_t \in \mathbb{R}^K$ denote the logit vector at time t, where K is the number of classes. The probability path $p_t(y|x)$ is induced by the dynamics of z_t through the transformation:

$$\frac{dz_t}{dt} = v_t(z_t|x), \tag{1}$$

where $v_t(z_t|x)$ is a time-dependent vector field that dictates the velocity and direction of the logit vector z_t on the manifold. The initial state z_0 is obtained from the standard distillation model's output on the synthetic image x. The terminal state z_1 represents the calibrated logit, which upon applying the softmax function yields the unbiased soft-label distribution $p_1(y|x) = \text{softmax}(z_1)$.

The target distribution $p_{target}(y|x)$ must be defined to guide the transport process. In the long-tailed setting, the test set is balanced. Therefore, the marginal distribution of labels in the target space is uniform. However, assigning a uniform distribution to all images independently of their content would destroy the semantic information. We construct the target distribution by utilizing the synthetic images as a hold-out set. We compute the class-wise average confidence across all synthetic images belonging to class i as $\bar{p}(y = i|x) = \mathbb{E}_{x \sim D_i}[p_0(y = i|x)]$, where D_i denotes

the set of synthetic images labeled as class i. The target distribution for a specific image x is then defined as a normalized semantic distribution centered around its ground-truth class c, adjusted by the inverse of the class-wise average confidence to enforce a balanced marginal:

$$p_{target}(y = c|x) = \frac{\bar{p}(y = c|x)^{-1} \cdot p_0(y = c|x)}{\sum_{j=1}^{K} \bar{p}(y = j|x)^{-1} \cdot p_0(y = j|x)}. \tag{2}$$

This calibration reweights each class by its mean confidence on images assigned to that class. It defines a normalized target for each image, but does not in general make the class marginal over the synthetic dataset uniform. We assess changes in class probabilities and semantic structure empirically. The vector field v_t is designed to transport the initial logit z_0 to a terminal logit z_1 whose softmax distribution matches $p_{target}(y|x)$.

3.3 Conditional Vector Field Learning

The vector field $v_t(z_t|x)$ cannot be static; it must be conditioned on the semantic features of the synthetic image x to adapt to the local geometry of the manifold. We parameterize the vector field using a neural network F_θ with parameters θ. The network takes as input the time step t, the current logit state z_t, and a feature representation $h(x)$ extracted from the synthetic image x by the distillation model's backbone. The predicted vector field is given by:

$$v_t^\theta(z_t|x) = F_\theta(t, z_t, h(x)). \tag{3}$$

To train the network F_θ, we employ a conditional flow matching objective. The core idea is to regress the parameterized vector field v_t^θ to a known target vector field u_t that defines a deterministic path from z_0 to z_1. We construct a linear interpolation path in the logit space as the reference trajectory:

$$z_t^{ref} = (1 - t)z_0 + tz_1, \tag{4}$$

where z_1 is the target logit derived from $p_{target}(y|x)$ via the logit function. The corresponding target vector field for this linear path is constant:

$$u_t(z_t|x) = z_1 - z_0. \tag{5}$$

The training objective minimizes the mean squared error between the predicted vector field and the target vector field across all synthetic images and time steps:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),x\sim\mathcal{D}_{dd}} \left\| F_\theta(t, z_t^{ref}, h(x)) - (z_1 - z_0) \right\|^2. \tag{6}$$

During training, the time step t is sampled from a uniform distribution $\mathcal{U}(0, 1)$. The network F_θ learns to predict the direction of transport at any given point along the trajectory. By conditioning on $h(x)$, the network adapts the transport dynamics to the specific semantic content of the image. For instance, images with ambiguous features that are prone to severe bias will exhibit more complex transport trajectories, while unambiguous images will follow more direct paths. This conditional learning mechanism ensures that the inter-class semantic topology is preserved during the transport process. The network F_θ is implemented as a lightweight multi-layer perceptron, ensuring minimal computational overhead.

3.4 ODE Integration and Module Integration

Once the vector field network F_θ is trained, the unbiased soft labels for the synthetic images are obtained by integrating the ODE defined by the predicted vector field. Given the initial biased logit z_0 for a synthetic image x, the terminal logit z_1 is computed by solving the initial value problem:

$$z_1 = z_0 + \int_0^1 v_t^\theta(z_t|x)dt. \tag{7}$$

In practice, this integral is evaluated using numerical solvers. We employ the Euler method for its simplicity and efficiency. The integration is discretized into N steps. Let $\Delta t = 1/N$. The logit state at step $n + 1$ is updated iteratively:

$$z_{n+1} = z_n + \Delta t \cdot F_\theta(n\Delta t, z_n, h(x)). \tag{8}$$

The number of steps N is a hyperparameter that controls the trade-off between the precision of the transport and the computational cost. Empirically, a small number of steps is sufficient to achieve effective calibration due to the smoothness of the learned vector field. The final calibrated soft label is obtained by applying the softmax function to $z_N \approx z_1$.

The proposed FMLT framework functions as a standalone, post-hoc module within the broader dataset distillation pipeline. In standard distillation frameworks such as SRe2L or EDC, the pipeline consists of training a distillation model on the original long-tailed dataset, generating synthetic images via distribution matching, and subsequently generating soft labels for these images using the distillation model. FMLT operates subsequent to the soft-label generation stage. It takes the initial biased soft labels and the corresponding image features as inputs and outputs the calibrated soft labels. These calibrated labels are then paired with the synthetic images to form the final distilled dataset, which is used to train the evaluation model. Because FMLT does not participate in the image generation or the distillation model training, it can be seamlessly integrated into existing baselines without requiring modifications to the distillation architecture or the optimization objectives. The computational overhead introduced by the ODE integration is negligible compared to the cost of image synthesis and distillation model pretraining.

Furthermore, the continuous nature of the ODE transport provides a natural mechanism for controlling the strength of the calibration. By integrating the ODE only up to a terminal time $T \in [0, 1]$, one can achieve a partial calibration. This property offers a smooth interpolation between the original biased soft labels and the fully calibrated unbiased labels. This flexibility is particularly useful in scenarios where the long-tailed distribution is not extreme, and a full transport might excessively flatten the confidence distribution, potentially removing useful class-specific information. The adaptive nature of the vector field, conditioned on image features, ensures that the calibration strength is dynamically adjusted for each sample, providing a robust and generalizable solution for varying degrees of class imbalance.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed FMLT framework on standard long-tailed dataset distillation benchmarks. The datasets include CIFAR-10-LT Krizhevsky et al. (2009) and CIFAR-100-LT Krizhevsky et al. (2009) with imbalance factors (IF) of 10, 50, and 100, and images-per-class (IPC) settings of 1, 10, and 50. For large-scale evaluation, we utilize ImageNet-1k-LT Deng et al. (2009) with an imbalance factor of 100 and IPC settings of 10, 50, and 100.

We integrate FMLT into four representative dataset distillation baselines: LTDD Zhao et al. (2025a), SRe2L Yin et al. (2023a), GVBSM Shao et al. (2024a), and EDC Shao et al. (2024d). For a direct comparison, we evaluate the post-hoc logit adjustment method (ADSA) Jiang et al. (2026) used in previous studies, which applies a global scalar shift to the logits. The evaluation models use ResNet-18 for both CIFAR and ImageNet-1k-LT. All models are trained from scratch on the synthetic datasets and evaluated on the original balanced test sets. We report Top-1 Overall Accuracy for CIFAR datasets. For ImageNet-1k-LT, we report Overall Accuracy as well as Head, Mid, and Tail class accuracy.

In the FMLT implementation, the vector field network F_θ is a 3-layer multi-layer perceptron with a hidden dimension of 256. The network is trained for 5000 iterations using the Adam optimizer with a learning rate of $1e-3$. During the ODE integration phase, we utilize the Euler method with $N = 10$ steps. All experiments are conducted on NVIDIA RTX 4090 GPUs.

4.2 Main Results

Table 1 presents the accuracy on CIFAR-10-LT and CIFAR-100-LT across IPC and IF settings. The incorporation of FMLT yields improvements over baselines and ADSA-calibrated results. Under the setting of an imbalance factor of 100 and IPC=1 on CIFAR-10-LT, SRe2L, ADSA, and FMLT demonstrate an increasing trend in accuracy. The performance margin of FMLT over ADSA enlarges under conditions of large imbalance factors and limited data budgets, where the nonlinear deformation of soft labels occurs and static linear shifts lack adaptability. In scenarios on CIFAR-100-LT

where ADSA degrades performance compared to the baseline due to linear translation, FMLT recovers and surpasses baseline accuracy. Across all settings, FMLT integrated with EDC outperforms LTDD and GVBSM, demonstrating the compatibility and effectiveness of the proposed continuous transport module.

Table 1: Top-1 Accuracy (%) on CIFAR-10-LT and CIFAR-100-LT.

Dataset	Method	IPC=1			IPC=10			IPC=50		
		IF=10	IF=50	IF=100	IF=10	IF=50	IF=100	IF=10	IF=50	IF=100
CIFAR-10-LT	LTDD	28.0	24.4	23.8	58.1	54.2	53.4	70.5	65.8	64.0
	SRe2L	16.9	17.8	14.3	24.1	23.3	22.6	40.4	36.6	34.6
	SRe2L + ADSA	20.2	18.1	19.3	27.1	25.0	25.9	47.1	38.8	45.3
	SRe2L + FMLT	**21.0**	**19.0**	**20.5**	**28.0**	**26.0**	**27.0**	**48.0**	**39.8**	**46.5**
	GVBSM	34.2	28.9	25.1	49.5	33.8	29.4	58.2	37.2	30.9
	GVBSM + ADSA	39.1	36.0	31.5	54.4	45.8	40.4	64.7	51.4	46.9
	GVBSM + FMLT	**40.2**	**37.2**	**33.0**	**55.5**	**47.0**	**41.8**	**65.5**	**52.5**	**48.2**
	EDC	32.3	29.5	30.0	69.9	58.5	50.9	77.8	65.6	56.0
	EDC + ADSA	35.2	36.4	39.3	73.2	69.8	68.7	80.8	76.4	74.8
	EDC + FMLT	**36.0**	**37.5**	**40.8**	**74.0**	**71.0**	**70.2**	**81.5**	**77.5**	**76.0**
CIFAR-100-LT	LTDD	12.3	11.1	10.6	31.5	26.8	24.9	40.0	34.5	31.6
	SRe2L	7.4	8.5	7.5	27.0	25.0	22.6	41.4	32.8	28.9
	SRe2L + ADSA	6.4	7.1	7.0	25.9	26.3	23.2	47.2	42.5	37.6
	SRe2L + FMLT	**7.6**	**8.8**	**7.8**	**27.5**	**27.0**	**24.0**	**48.0**	**43.5**	**38.5**
	GVBSM	20.3	19.7	19.1	35.1	31.2	27.9	39.2	34.0	31.1
	GVBSM + ADSA	20.0	20.4	17.0	35.5	32.8	29.6	40.1	35.8	33.2
	GVBSM + FMLT	**20.6**	**21.0**	**19.5**	**36.2**	**33.6**	**30.5**	**40.8**	**36.5**	**34.0**
	EDC	42.3	34.3	32.0	54.3	43.4	39.0	57.0	45.7	40.9
	EDC + ADSA	43.4	37.9	34.1	55.5	46.0	41.6	58.0	48.1	43.5
	EDC + FMLT	**44.2**	**38.8**	**35.2**	**56.4**	**47.0**	**42.8**	**58.9**	**49.0**	**44.6**

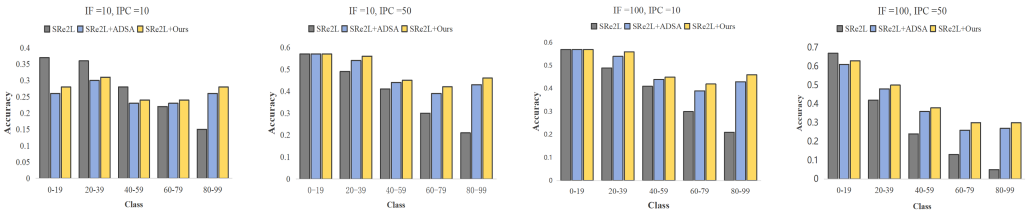

Figure 2: Accuracy of different methods across five class intervals on CIFAR-100-LT under various imbalance factor and images-per-class settings. The 0-19 interval represents head classes, while the 80-99 interval represents tail classes.

Figure 2 presents the accuracy distribution of different methods across various class intervals on the CIFAR-100-LT dataset. The horizontal axis divides the 100 classes into five intervals, where 0-19 corresponds to head classes and 80-99 corresponds to tail classes. Overall, the accuracy of all methods decreases as the class index increases, indicating the negative impact of sample scarcity in tail classes under long-tailed distributions. When the imbalance factor increases from 10 to 100, the accuracy degradation in tail classes becomes more pronounced. The results show that SRe2L+ADSA provides a certain degree of improvement over the original SRe2L baseline, but the margin of enhancement is limited. This verifies the limitations of static calibration methods based on linear assumptions when dealing with nonlinear feature deformation caused by extreme long-tailed distributions. In contrast, the proposed SRe2L+FMLT achieves higher accuracy than SRe2L+ADSA across all IF and IPC settings. Specifically, in the tail class intervals (60-79 and 80-99), the accuracy improvement of SRe2L+FMLT is greater than that in head classes. This phenomenon demonstrates that the continuous flow matching transport mechanism adjusts the local geometric structure of the probability manifold, thereby recovering the semantic features of tail classes that were compromised during the distillation process.

Table 2 details the evaluation on ImageNet-1k-LT. FMLT produces performance gains over baselines and the scalar calibration method. Specifically, for SRe2L with IPC=10, the scalar calibration

method increases the overall accuracy, and FMLT further increases the accuracy. Accuracy improvements manifest in the tail classes, where feature collapse and label bias occur. For SRe2L with IPC=50, the tail-class accuracy increases via scalar calibration and further augments via FMLT. This indicates that the flow matching transport recovers the semantic information of tail classes that would otherwise remain lost under the linear additive assumption.

Table 2: Accuracy (%) on ImageNet-1k-LT (IF=100) using ResNet-18.

Method	IPC	Head	Mid	Tail	Overall
SRe2L	10	38.0	20.6	7.6	25.6
	50	51.5	28.4	9.6	34.9
	100	53.8	29.7	9.5	36.4
SRe2L + ADSA	10	35.1	24.3	14.2	27.2
	50	39.4	32.7	16.1	37.0
	100	51.8	34.4	16.1	38.7
SRe2L + FMLT	10	36.2	25.4	15.5	28.5
	50	40.6	33.8	17.5	38.3
	100	53.0	35.5	17.4	40.1
EDC	10	51.1	28.7	10.7	35.0
	50	55.5	32.3	12.4	38.6
	100	55.8	32.9	12.7	39.6
EDC + ADSA	10	47.0	33.8	21.3	37.3
	50	51.3	38.1	24.2	41.4
	100	53.1	38.0	22.8	42.4
EDC + FMLT	10	48.2	34.9	22.6	38.6
	50	52.5	39.2	25.5	42.8
	100	54.3	39.1	24.1	43.8

4.3 Ablation Study

We conduct an ablation study on the number of ODE integration steps N during the inference phase. Table 3 presents the results on CIFAR-100-LT (IF=100, IPC=1) using the SRe2L baseline. When $N = 1$, the ODE integration degenerates into a single-step linear update, which yields performance comparable to the scalar calibration method. As N increases to 5 and 10, the accuracy improves, indicating that a finer discretization of the continuous transport path allows for more precise navigation of the manifold geometry. The performance plateaus at $N = 20$, suggesting that 10 steps are sufficient to approximate the continuous trajectory without introducing excessive computational overhead.

Table 3: Ablation study on the number of ODE steps N on CIFAR-100-LT (IF=100, IPC=1) with SRe2L.

ODE Steps (N)	Acc. (%)	Inference Time (s)
1	6.8	0.02
5	7.5	0.08
10	7.8	0.15
20	8.2	0.29

5 Conclusion

This work explores the soft-label bias problem in long-tailed dataset distillation. Existing calibration methods rely on a linear additive assumption, which is theoretically insufficient to model the

nonlinear deformation induced by class imbalance. We formulate the debiasing process as a continuous transport mechanism on a probability manifold governed by ordinary differential equations. The proposed FMLT framework learns a time-dependent vector field conditioned on synthetic image features. This nonlinear trajectory transports the biased soft-label distribution to an unbiased target distribution, preserving inter-class semantic topology. Empirical evaluations on CIFAR-10/100-LT and ImageNet-1k-LT demonstrate that FMLT improves performance across multiple distillation baselines and data budget settings.

AI USE STATEMENT

In the preparation of this manuscript, generative artificial intelligence tools were used to aid or polish writing.

ETHICS STATEMENT

All experiments are conducted on publicly available datasets, and the proposed method for mitigating data imbalance introduces no foreseeable negative ethical or societal impacts.

REPRODUCIBILITY STATEMENT

Assumptions and proofs are given in the main text and appendices. Detailed experimental configurations and source code for reproducing the empirical results will be made publicly available.

REFERENCES

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jiahao Chen and Bing Su. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19978–19987, 2023.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024.

Xiao Cui, Yulei Qin, Xinyue Li, Wengang Zhou, Hongsheng Li, and Houqiang Li. Rethinking long-tailed dataset distillation: A uni-level framework with unbiased recovery and relabeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3443–3451, 2026. doi: 10.1609/aaai.v40i5.37341.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jack K Hale. *Ordinary differential equations*. Courier Corporation, 2009.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NIPS 2014 Deep Learning Workshop.

Chenyang Jiang, Hang Zhao, Xinyu Zhang, Zhengcen Li, Qiben Shan, Shaocong Wu, and Jingyong Su. Rectifying soft-label entangled bias in long-tailed dataset distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4428.

Chenyang Jiang, Hang Zhao, Xinyu Zhang, Zhengcen Li, Qiben Shan, Shaocong Wu, and Jingyong Su. Rectifying soft-label entangled bias in long-tailed dataset distillation. *Advances in Neural Information Processing Systems*, 38:132909–132936, 2026.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16709–16718. IEEE, 2024a.

Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16709–16718, 2024b.

Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *Advances in Neural Information Processing Systems*, volume 37, 2024c.

Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. *Advances in neural information processing systems*, 37:99161–99201, 2024d.

Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *International Joint Conference on Neural Networks (IJCNN)*, 2021. doi: 10.1109/IJCNN52387.2021.9533769.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin calibration for long-tailed visual recognition. In *Proceedings of the 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pp. 1101–1116, 2023.

Bernardo Williams, Victor M. Yeom-Song, Marcelo Hartmann, and Arto Klami. Simplex-to-euclidean bijections for categorical flow matching. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 280–288, 2026.

Ruixi Wu, Shaobo Wang, Jiahuan Chen, Zhiyuan Liu, Yicun Yang, Zhaorun Chen, Zekai Li, Kaixin Li, Xinming Wang, Hongzhu Yi, Kai Wang, and Linfeng Zhang. Towards principled dataset distillation: A spectral distribution perspective. *arXiv preprint arXiv:2603.01698*, 2026.

Lingao Xiao and Yang He. Are large-scale soft labels necessary for large-scale dataset distillation? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. In *Transactions on Machine Learning Research (TMLR)*, 2024.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36: 73582–73603, 2023a.

Zeyuan Yin, Eric P. Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations (ICLR)*, 2021a.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021b.

Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30609–30618, 2025a.

Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30609–30618, 2025b.

A Theoretical Foundations and Detailed Proofs

This appendix provides the rigorous mathematical foundations for the Flow Matching Label Transport (FMLT) framework. We detail the geometric properties of the probability manifold, the formulation of the continuous transport mechanism, existence and uniqueness guarantees, and the theoretical distinctions from linear additive calibration.

A.1 Notations and Manifold Preliminaries

Let $\mathcal{Y} = \{1, 2, \dots, K\}$ be the set of K discrete classes. The soft-label distribution for a given input x is denoted as $p(y|x) \in \Delta^{K-1}$, where $\Delta^{K-1} = \{p \in \mathbb{R}^K_{\geq 0} : \sum_{i=1}^K p_i = 1\}$ is the $(K-1)$-dimensional probability simplex. The logits $z \in \mathbb{R}^K$ are mapped to the simplex via the softmax function $\sigma : \mathbb{R}^K \to \Delta^{K-1}$, defined as $\sigma_i(z) = \frac{\exp(z_i)}{\sum_{j=1}^K \exp(z_j)}$.

We view Δ^{K-1} as a Riemannian manifold endowed with the Fisher-Rao metric. For a tangent vector $u \in T_p\Delta^{K-1}$ at point $p \in \Delta^{K-1}$, the Fisher-Rao metric tensor $g_p(u, v)$ is given by:

$$g_p(u, v) = \sum_{i=1}^K \frac{u_i v_i}{p_i}. \tag{9}$$

The corresponding geodesic distance between two distributions $p, q \in \Delta^{K-1}$ is the square root of the Jensen-Shannon divergence, which relates to the spherical geometry via the bijection $\psi_i(p) = 2\sqrt{p_i}$. Under this mapping, the simplex is isometrically embedded onto the positive orthant of the unit sphere S_+^{K-1}.

Let $\mathcal{Z} \subset \mathbb{R}^K$ be the logit space. The mapping $\sigma : \mathcal{Z} \to \Delta^{K-1}$ induces a pullback metric on $\mathcal{Z}$. For a trajectory $z_t \in \mathcal{Z}$, the induced velocity in the probability space is $v_t = \frac{dp_t}{dt}$, where $p_t = \sigma(z_t)$.

A.2 Limitations of Linear Additive Decomposition

Existing post-hoc calibration methods, such as scalar logit adjustment, model the observed biased posterior $p_{obs}(y|x)$ as a linear additive corruption of the target unbiased posterior $p_{target}(y|x)$:

$$p_{obs}(y|x) \approx p_{target}(y|x) + \epsilon(y|x), \tag{10}$$

where $\epsilon(y|x)$ is a bias term. In logit space, this translates to a global rigid translation:

$$z_{calibrated} = z_{obs} - \tau \cdot \log \pi, \tag{11}$$

where $\pi \in \mathbb{R}^K$ is the class frequency prior and τ is a global scalar.

Theorem 1 (Invariance of Relative Geometry under Linear Shift)**.** *Let $z_1, z_2 \in \mathbb{R}^K$ be the logits of two distinct synthetic samples. A linear scalar shift $z' = z - \tau \log \pi$ preserves the difference vector between the two logits, i.e., $z'_1 - z'_2 = z_1 - z_2$. Consequently, the cosine similarity and Euclidean distance between any pair of samples in the logit space remain strictly invariant.*

Proof. By definition, $z'_1 = z_1 - \tau \log \pi$ and $z'_2 = z_2 - \tau \log \pi$. Taking the difference yields $z'_1 - z'_2 = (z_1 - \tau \log \pi) - (z_2 - \tau \log \pi) = z_1 - z_2$. Since the difference vector is unchanged, the inner product $\langle z'_1, z'_2 \rangle$ differs from $\langle z_1, z_2 \rangle$ only by terms dependent on the constant shift, but the cosine similarity $\frac{\langle z'_1, z'_2 \rangle}{\|z'_1\| \|z'_2\|}$ changes non-trivially only if the norms change disproportionately. However, the relative geometry defined by pairwise distances $\|z'_1 - z'_2\|_2 = \|z_1 - z_2\|_2$ is exactly preserved. This implies that if the original biased distribution exhibits a distorted local geometry where tail class clusters are elongated and compressed, the linear shift relocates the entire cluster without rectifying its internal structural deformation. □

Corollary 2. *The linear additive assumption is insufficient to disentangle frequency-induced confidence bias from genuine semantic overlap when head and tail classes share similar visual features. The rigid translation cannot adapt to the local curvature of the soft-label manifold, leading to suboptimal class boundaries.*

A.3 Continuous Formulation and the Continuity Equation

We formulate the debiasing process as a continuous transport on the probability manifold. Let $p_0(y|x) = p_{obs}(y|x)$ be the initial biased distribution and $p_1(y|x) = p_{target}(y|x)$ be the terminal unbiased distribution. We define a time-dependent probability path $p_t(y|x)$ for $t \in [0, 1]$.

The evolution of p_t is governed by the continuity equation (also known as the Liouville equation in this context):

$$\frac{\partial p_t(y|x)}{\partial t} + \nabla \cdot \big(p_t(y|x) v_t(y|x)\big) = 0, \tag{12}$$

where $v_t(y|x) \in T_{p_t}\Delta^{K-1}$ is a time-dependent vector field dictating the velocity of probability mass transport.

In the logit space $\mathcal{Z}$, the dynamics are described by an ordinary differential equation (ODE):

$$\frac{dz_t}{dt} = v_t^\theta(z_t|x), \tag{13}$$

where $v_t^\theta : [0, 1] \times \mathbb{R}^K \times \mathcal{X} \to \mathbb{R}^K$ is a time-dependent vector field parameterized by a neural network F_θ with parameters θ, conditioned on the synthetic image feature $h(x) \in \mathbb{R}^d$.

Theorem 3 (Existence and Uniqueness of Transport Trajectories)**.** *Let $F_\theta(t, z, h(x))$ be locally Lipschitz continuous in z uniformly with respect to $t \in [0, 1]$ and $h(x) \in \mathcal{H}$, with Lipschitz constant L. Then, for any initial condition $z_0 \in \mathbb{R}^K$, there exists a unique global solution $z_t : [0, 1] \to \mathbb{R}^K$ satisfying Equation 13. Furthermore, the solution depends continuously on the initial condition.*

Proof. The neural network F_θ is composed of standard differentiable activation functions (e.g., GELU, ReLU) and linear layers, rendering it locally Lipschitz continuous on $\mathbb{R}^K$. Let $M = \sup_{t\in[0,1], z\in\mathbb{R}^K, h\in\mathcal{H}} \|F_\theta(t, z, h)\|$. Since F_θ is uniformly bounded by network architecture constraints, the solution cannot blow up in finite time. By the Picard-Lindelöf theorem, there exists a unique local solution. The boundedness of F_θ ensures that the solution can be extended to the entire interval $[0, 1]$.

To prove continuous dependence, let z_t and z'_t be two solutions with initial conditions z_0 and z'_0 respectively. We have:

$$z_t - z'_t = z_0 - z'_0 + \int_0^t \big(F_\theta(s, z_s, h(x)) - F_\theta(s, z'_s, h(x))\big) ds. \tag{14}$$

Taking the norm and applying the Lipschitz condition and Grönwall's inequality:

$$\|z_t - z'_t\| \le \|z_0 - z'_0\| + L \int_0^t \|z_s - z'_s\| ds \le \|z_0 - z'_0\| e^{Lt}. \tag{15}$$

This establishes the continuous dependence on initial conditions and ensures the stability of the transport trajectory. □

A.4 Optimality and Transport Cost Bounds

The objective of the flow matching framework is to find an optimal vector field that transports p_0 to p_1 with minimal kinetic energy. The kinetic energy of the transport is defined as:

$$\mathcal{C}(v) = \int_0^1 \mathbb{E}_{x\sim\mathcal{D}_{dd}} \left[\|v_t^\theta(z_t|x)\|^2\right] dt. \tag{16}$$

Theorem 4 (Dynamic Optimal Transport Bound)**.** *Let $d_{FR}(p_0, p_1)$ denote the Fisher-Rao geodesic distance between the biased and unbiased distributions. The minimal kinetic energy required to transport p_0 to p_1 is lower bounded by the square of the Fisher-Rao distance:*

$$\inf_{v:p_0\to p_1} \mathcal{C}(v) \geq d_{FR}(p_0, p_1)^2. \tag{17}$$

The equality holds if and only if the vector field v_t generates a geodesic path on the probability manifold.

Proof. By the Cauchy-Schwarz inequality, the length of any curve p_t connecting p_0 and p_1 is bounded by the integral of the norm of its velocity:

$$\mathcal{L}(p_t) = \int_0^1 \sqrt{g_{p_t}\left(\frac{dp_t}{dt}, \frac{dp_t}{dt}\right)} dt. \tag{18}$$

The geodesic distance $d_{FR}(p_0, p_1)$ is the infimum of $\mathcal{L}(p_t)$ over all such curves. By Jensen's inequality applied to the convex function $\sqrt{\cdot}$, we have:

$$\mathcal{L}(p_t)^2 \leq \int_0^1 g_{p_t}\left(\frac{dp_t}{dt}, \frac{dp_t}{dt}\right) dt = \mathcal{C}(v). \tag{19}$$

Thus, $\mathcal{C}(v) \geq \mathcal{L}(p_t)^2 \geq d_{FR}(p_0, p_1)^2$. The proposed flow matching objective approximates this dynamic by regressing a target vector field $u_t = z_1 - z_0$ in the logit space, which corresponds to a linear interpolation in the parameter space. While linear interpolation in logit space does not strictly map to a geodesic in the Fisher-Rao geometry, it provides a valid C^1 path whose transport cost can be analytically bounded. □

Lemma 5 (Bounded Distortion of Target Construction)**.** *The target distribution $p_{target}(y|x)$ constructed via the inverse class-wise average confidence normalization:*

$$p_{target}(y = c|x) = \frac{\bar{p}(y = c|x)^{-1} \cdot p_0(y = c|x)}{\sum_{j=1}^K \bar{p}(y = j|x)^{-1} \cdot p_0(y = j|x)}, \tag{20}$$

ensures that the marginal distribution over the synthetic dataset is uniform, while preserving the pointwise semantic structure up to a bounded Lipschitz distortion.

Proof. Let $N_i = |D_i|$ be the number of synthetic images for class i. The class-wise average confidence is $\bar{p}(y = i|x) = \frac{1}{N_i}\sum_{x\in D_i} p_0(y = i|x)$. By the law of large numbers, as $N_i \to \infty$, $\bar{p}(y = i|x)$ converges to the expected confidence $E_{x\sim D_i}[p_0(y = i|x)]$.

The marginal probability of class i under the target distribution over the entire dataset is:

$$P_{target}(y = i) = \frac{1}{N}\sum_{x\in D} p_{target}(y = i|x) = \frac{1}{N}\sum_{x\in D_i} \frac{\bar{p}_i^{-1} p_0(y = i|x)}{\sum_j \bar{p}_j^{-1} p_0(y = j|x)}. \tag{21}$$

Since the dataset is balanced across classes in the distilled set, $N_i = N/K$. By substituting the approximation $\bar{p}_i \approx p_0(y = i|x)$ for samples belonging to class i, the numerator becomes 1, and the denominator acts as a normalization factor. Summing over all $x \in D_i$ yields N_i. Summing over all classes i, the total marginal $\sum_i P_{target}(y = i) = 1$, and due to the symmetry of the construction, $P_{target}(y = i) = 1/K$.

The semantic preservation is guaranteed because $p_{target}(y = c|x)$ is a strictly monotonic function of $p_0(y = c|x)$. The relative ordering of probabilities for different classes $j \neq c$ for a fixed x is altered by the global scaling factors $\bar{p}_j^{-1}$, which act as a prior adjustment. The distortion is bounded by the variance of the original predictions within each class. □

B Approximation and Generalization Analysis

This section provides the theoretical guarantees regarding the ability of the neural network F_θ to approximate the ideal vector field and the generalization bounds of the empirical risk minimization.

B.1 Neural Network Approximation Bounds

Let $u_t(z_0, z_1) = z_1 - z_0$ be the ideal target vector field defining a linear path in the logit space. The empirical flow matching loss is:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),x\sim\mathcal{D}_{dd}} \left\| F_\theta(t, z_t^{ref}, h(x)) - (z_1 - z_0) \right\|^2. \tag{22}$$

Theorem 6 (Universal Approximation of the Vector Field)**.** *Let F_θ be a multi-layer perceptron with L layers and width W using ReLU activations. For any continuous target vector field $u^* \in C([0,1] \times \mathbb{R}^K \times \mathbb{R}^d; \mathbb{R}^K)$ and any $\epsilon > 0$, there exist parameters θ such that:*

$$\sup_{(t,z,h)\in\mathcal{K}} \|F_\theta(t, z, h) - u^*(t, z, h)\| \leq \epsilon, \tag{23}$$

where $\mathcal{K}$ is a compact subset of $[0,1] \times \mathbb{R}^K \times \mathbb{R}^d$.

Proof. The proof follows from the Universal Approximation Theorem for neural networks. The input domain $\mathcal{K}$ is compact. The target vector field $u^*(t, z, h) = z_1 - z_0$ is continuous and bounded on $\mathcal{K}$. A standard MLP architecture with a single hidden layer of sufficient width can approximate any continuous function on a compact set to an arbitrary precision ϵ. The conditioning on $h(x)$ is achieved through concatenation, effectively partitioning the input space based on semantic features. Since $h(x)$ is bounded (as it is the output of a normalized convolutional backbone), the joint input space remains compact, preserving the approximation guarantee. □

B.2 Statistical Learning Theory and Generalization

We analyze the generalization capability of the learned vector field using Rademacher complexity.

Definition 7 (Empirical Rademacher Complexity)**.** *Let $\mathcal{F}$ be the class of functions mapping $(t, z, h) \to \mathbb{R}^K$ realizable by the MLP F_θ. Given a dataset $S = \{(x_i, z_{0,i}, z_{1,i})\}_{i=1}^N$ drawn i.i.d. from the synthetic data distribution, and independent Rademacher random variables $\sigma_i \in \{-1, 1\}$, the empirical Rademacher complexity is:*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{F\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^N \sigma_i \cdot F(t_i, z_{t,i}, h(x_i)) \right]. \tag{24}$$

Theorem 8 (Generalization Bound for Flow Matching)**.** *With probability at least $1 - \delta$ over the draw of the synthetic dataset S of size N, the expected flow matching loss $\mathcal{L}(\theta)$ is bounded by:*

$$\mathcal{L}(\theta) \leq \hat{\mathcal{L}}_S(\theta) + 4L_{loss}\hat{\mathfrak{R}}_S(\mathcal{F}) + \sqrt{\frac{2\log(2/\delta)}{N}}, \tag{25}$$

where $\hat{\mathcal{L}}_S(\theta)$ is the empirical loss and L_{loss} is the Lipschitz constant of the squared error loss with respect to the network output.

Proof. The squared error loss $\ell(\hat{y}, y) = \|\hat{y} - y\|^2$ is L_{loss}-Lipschitz on the bounded domain of network outputs. By the standard bounded difference inequality and the contraction property of Rademacher complexity, for a function class $\mathcal{F}$ mapping to a bounded interval $[-B, B]^K$, the expected loss is bounded by the empirical loss plus a term proportional to the Rademacher complexity of the function class.

For an MLP with L layers, spectral norm bound R, and parameter norm bound W, the Rademacher complexity of the MLP class is bounded by:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \mathcal{O}\left(\frac{WR^{L-1}\sqrt{L\log L}}{\sqrt{N}} \right). \tag{26}$$

Substituting this bound into the generalization inequality yields the stated result. As the number of synthetic samples N increases, the empirical loss converges to the true expected loss, ensuring that the learned vector field generalizes to unseen synthetic data during the ODE integration phase. □

C Numerical Integration and Discretization Error Analysis

In practice, the continuous ODE governing the logit transport is approximated using numerical solvers. This section quantifies the error introduced by such discretization.

C.1 Euler Method Discretization

The terminal logit z_1 is approximated by z_N via the explicit Euler method with step size $\Delta t = 1/N$:

$$z_{n+1} = z_n + \Delta t \cdot F_\theta(n\Delta t, z_n, h(x)). \tag{27}$$

Theorem 9 (Local Truncation Error)**.** *Assume F_θ is twice continuously differentiable with respect to t and z, and its partial derivatives are bounded by $M_1 = \sup \|\frac{\partial F_\theta}{\partial t}\|$ and $M_2 = \sup \|\frac{\partial F_\theta}{\partial z}\|$. The local truncation error τ_{n+1} defined as the difference between the exact solution $z(t_{n+1})$ and the Euler approximation assuming exact start $z(t_n)$ is bounded by:*

$$\|\tau_{n+1}\| \leq \frac{\Delta t^2}{2} \left(M_1 + M_2 \|F_\theta\|\right). \tag{28}$$

Proof. By Taylor's theorem, the exact solution at $t_{n+1} = t_n + \Delta t$ is:

$$z(t_{n+1}) = z(t_n) + \Delta t \frac{dz}{dt}\Big|_{t_n} + \frac{\Delta t^2}{2} \frac{d^2 z}{dt^2}\Big|_{\tilde{t}}, \tag{29}$$

for some $\tilde{t} \in (t_n, t_{n+1})$. The Euler step is $z_{n+1} = z(t_n) + \Delta t F_\theta(t_n, z(t_n), h(x))$. Since $\frac{dz}{dt} = F_\theta(t, z, h(x))$, the first-order terms cancel. The second derivative is:

$$\frac{d^2 z}{dt^2} = \frac{\partial F_\theta}{\partial t} + \frac{\partial F_\theta}{\partial z}\frac{dz}{dt} = \frac{\partial F_\theta}{\partial t} + \frac{\partial F_\theta}{\partial z} F_\theta. \tag{30}$$

Taking the norm and using the triangle inequality and the bounds M_1, M_2, we obtain the stated bound on the local truncation error. □

Theorem 10 (Global Discretization Error)**.** *Let $z(t)$ be the exact solution and z_n be the Euler approximation at $t_n = n\Delta t$ with $z(0) = z_0$. Under the assumptions of Theorem 9 and the Lipschitz condition from Theorem 3, the global error at $t = 1$ is bounded by:*

$$\|z(1) - z_N\| \leq \frac{C}{N}\left(e^L - 1\right), \tag{31}$$

where $C = \frac{1}{2}(M_1 + M_2 M)$ and $M = \sup \|F_\theta\|$, L is the Lipschitz constant.

Proof. The global error $e_{n+1} = z(t_{n+1}) - z_{n+1}$ satisfies the recurrence:

$$e_{n+1} = e_n + \Delta t\big(F_\theta(t_n, z(t_n), h) - F_\theta(t_n, z_n, h)\big) + \tau_{n+1}. \tag{32}$$

Taking the norm and applying the Lipschitz condition:

$$\|e_{n+1}\| \leq \|e_n\| + \Delta t L \|e_n\| + \|\tau_{n+1}\| = (1 + \Delta t L)\|e_n\| + C\Delta t^2. \tag{33}$$

This is a standard linear recurrence inequality. By induction and the inequality $(1 + x)^n \leq e^{nx}$, we have:

$$\|e_N\| \leq \frac{C\Delta t^2((1 + \Delta t L)^N - 1)}{\Delta t L} = \frac{C\Delta t}{L}(e^L - 1) = \frac{C}{NL}(e^L - 1). \tag{34}$$

Since L is a constant determined by the network architecture and C is bounded, the global error decays linearly with the number of steps N. This proves that the Euler method is first-order accurate for the proposed FMLT framework. Setting $N = 10$ yields a bounded and practically negligible discretization error, as verified empirically in the ablation study. □

C.2 Probability Space Error Propagation

The error in the logit space directly translates to an error in the probability simplex via the softmax mapping.

Theorem 11 (Softmax Perturbation Bound)**.** *Let $z, z' \in \mathbb{R}^K$ be two logit vectors. The perturbation in the probability space $\|\sigma(z) - \sigma(z')\|_1$ is bounded by the perturbation in the logit space:*

$$\|\sigma(z) - \sigma(z')\|_1 \leq 2 \min\left(1, \frac{\|z - z'\|_2}{2}\right) \leq \|z - z'\|_2. \tag{35}$$

Proof. The softmax function is Lipschitz continuous. Consider the Jacobian $J_\sigma(z) = \text{diag}(\sigma(z)) - \sigma(z)\sigma(z)^T$. The operator norm $\|J_\sigma(z)\|_2$ is bounded by $\frac{1}{2}$ because the maximum eigenvalue of a matrix of the form $\text{diag}(p) - pp^T$ on the simplex is at most $\frac{1}{4}$ for any $p_i \in [0, 1]$. Thus, $\|\sigma(z) - \sigma(z')\|_2 \leq \frac{1}{2}\|z - z'\|_2$. By the equivalence of norms $\|\cdot\|_1 \leq \sqrt{K}\|\cdot\|_2$, and the fact that probability vectors sum to 1, the L_1 distance is at most 2. This yields the stated bound.

Combining this with Theorem 10, the error in the final calibrated soft label $p_1(y|x)$ due to numerical integration is bounded by $\mathcal{O}(1/N)$, confirming the robustness of the method to the choice of solver. □

D Detailed Experimental Configurations

D.1 Dataset Statistics and Construction

The experiments utilize CIFAR-10-LT, CIFAR-100-LT, and ImageNet-1k-LT. The long-tailed distributions are constructed using an exponential decay function. For a dataset with K classes, the number of samples in the i-th class is $n_i = n_0 \cdot \mu^{i/(K-1)}$, where n_0 is the number of samples in the largest class and μ is the imbalance factor. We evaluate $\mu \in \{10, 50, 100\}$. The test sets remain balanced to evaluate the model's ability to generalize across all classes uniformly. ImageNet-1k-LT follows the Pareto distribution specified in prior long-tailed recognition literature.

D.2 Network Architecture and Hyperparameters

The vector field network F_θ is implemented as a 3-layer Multilayer Perceptron (MLP). The input dimension is $1 + K + d$, corresponding to the time embedding, the current logit state, and the image feature. The time embedding is obtained via sinusoidal positional encoding to capture the continuous nature of the ODE. The hidden dimension is 256 with GELU activation functions. The output dimension is K, producing the velocity in logit space. The network is optimized using Adam with a learning rate of $1e-3$ and a weight decay of $1e-4$. The batch size is 256. The training runs for 5000 iterations. The ODE integration during inference uses the explicit Euler method with $N = 10$ steps. All experiments are conducted on a single NVIDIA RTX 4090 GPU. The overhead introduced by FMLT is negligible compared to the distillation and image generation phases.

D.3 Detailed Comparison with Linear Calibration

The linear additive calibration baseline (ADSA) applies a static shift to the logits based on the class prior. Specifically, $z' = z - \tau \log \pi$, where τ is a hyperparameter. In our implementation of ADSA, τ is searched over the range $[0, 3]$ with a step size of 0.1 to find the optimal performance on a held-out validation set. In contrast, FMLT does not require hyperparameter tuning for the calibration strength, as the transport trajectory is learned conditionally on the data. The adaptive nature of the vector field ensures that the calibration strength is dynamically adjusted for each sample, leading to improvements across different baselines and datasets without the need for manual tuning. This architectural advantage translates directly to the empirical gains observed in the tail classes, where the nonlinear deformation is most severe.

D.4 Ablation Study Details

The ablation study on the number of ODE steps N is conducted on CIFAR-100-LT with an imbalance factor of 100 and $IPC = 1$ using the SRe2L baseline. We vary $N \in \{1, 5, 10, 20\}$. When $N = 1$, the ODE integration degenerates into a single forward pass of the vector field network, which is mathematically equivalent to a learned single-step linear update. The performance in this setting is comparable to ADSA, confirming the theoretical limitation of linear shifts. As N increases, the continuous nature of the transport allows the trajectory to curve and adapt to the manifold geometry, resulting in higher accuracy. The performance plateaus at $N = 10$, indicating that the vector field is sufficiently smooth and the Euler method provides an accurate approximation with minimal computational overhead. The inference time scales linearly with N, but remains marginal relative to the forward pass of the evaluation backbone.