USEFUL BEFORE IT GENERATES WELL: MODEL-ALIGNED SPARSE SINKHORN FOR FLOW RETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Partially trained flow models are usually judged by sample quality, yet they may already carry enough coarse geometric structure to improve later training. We ask whether such early-stage models can be useful before they generate well, by serving as coupling constructors for flow retraining. The available coupling choices fail differently. Independent coupling is simple but high variance. Rectified-Flow-style model-generated coupling reduces variance but inherits teacher and solver drift. Optimal-transport couplings are principled but too expensive at dataset scale. MASS (Model-Aligned Sparse Sinkhorn) uses model-generated endpoints from an early-stage flow only to define candidate neighborhoods in the real dataset, then applies semi-relaxed Sinkhorn reweighting on the resulting sparse graph to produce a real-data-anchored transport plan at $O(kN)$ cost rather than $O(N^2)$. The plan is built once before retraining and reused across all training steps, unlike minibatch OT methods that re-solve an OT problem at every batch, and it composes with any flow-matching objective. We characterize sufficient conditions on the neighborhood size k under which the MASS empirical target marginal provably improves over the pretrained model's output distribution. On a partially trained CIFAR-10 teacher, the induced-target FID drops from 21.75 (k=1 NN) to 0.62 (k=200 MASS), and CIFAR-10 retraining FID improves from 21.43 (iMF) and 15.53 (Dec-iMF) to 12.59 (MASS-ReiMF) at matched compute. On 2D moons, MASS achieves transport cost within 3% of full global OT.

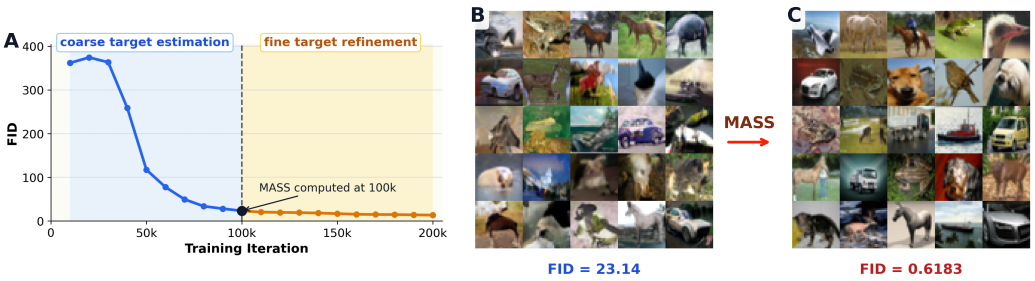

Figure 1: (A) FID drops rapidly first, then improves more slowly during refinement. (B) Samples from the 100k checkpoint still have poor fidelity. (C) MASS maps the same source-noise positions to transported real targets, producing a much lower-FID target set for retraining.

1 INTRODUCTION

In flow-based generative modeling, the learning signal depends on the model class and on how source noise is coupled to target data. Standard Flow Matching (Lipman et al., 2023) uses independent coupling, which is simple but often yields curved trajectories. One response is to retrain with model-generated coupling, as in Rectified Flow (Liu et al., 2023), where each source sample is paired with a sample produced by a pretrained model. Another response is to replace independent coupling with transport-aware pairing, for example in OT-based Flow Matching (Tong et al., 2024; Pooladian et al., 2023). These approaches all point in the same direction, that coupling quality matters, and they fail

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.

differently. Rectified-Flow-style coupling can inherit teacher error and numerical drift; global OT is too expensive; minibatch OT can be too local when the batch size is small.

Training curves suggest a useful intermediate regime. In many flow models, quality improves quickly early in training and slows later, which indicates that coarse target structure is learned before high-fidelity generation. This shifts the question. The issue is not whether a partially trained model is already a good generator, but whether it already contains enough transport-relevant structure to support a better coupling for the next stage of training. If so, we should be able to exploit that structure without treating model-generated samples as final targets, and without inheriting the drift that can accompany Rectified-Flow-style coupling.

We propose Model-Aligned Sparse Sinkhorn (MASS), a one-time coupling construction for flow retraining. MASS starts from model-generated endpoints of an early-stage model, but uses them only as geometric hints. For each endpoint, it retrieves a small neighborhood in the real dataset, then applies semi-relaxed Sinkhorn-style reweighting on the resulting sparse graph. The final plan is anchored to real data, computed once before retraining, and reused throughout training. Because MASS only changes how source samples are paired with real targets, it is a coupling primitive rather than a training method, and composes with standard FM, Mean Flow, iMF, and their rectified variants without modifying their objectives.

The construction also gives a natural bias and variance view of coupling. Independent pairing is unbiased but high variance. Rectified Flow coupling can have very low variance, but may be biased by model and solver drift. MASS interpolates between these extremes through model-guided neighborhoods and a tunable neighborhood size k. Small k keeps variance low but risks coverage and bias errors, while larger k improves coverage at higher computational cost. MASS is therefore not only a sparse OT approximation, but a practical way to control coupling quality in retraining.

Our main contributions are as follows.

- **A new framing for flow retraining.** We study whether a partially trained flow model can become useful before it becomes a good generator, by supplying enough coarse geometric signal to construct a better coupling for later training.
- **The MASS algorithm.** We introduce a model-guided sparse OT construction that uses model-generated endpoints to define candidate neighborhoods, then reweights them into a transport plan anchored to real data. MASS reduces coupling construction from $O(N^2)$ memory and time for full Sinkhorn to $O(kN)$ on the sparse graph, with much smaller k sufficient in practice.
- **Threshold-k certificates via two-measure concentration.** We give explicit sufficient conditions on k under which the MASS empirical target marginal provably improves over externally supplied baseline quality levels in both KL divergence and W_2 distance. The technical core is a two-measure k-NN degree concentration argument in which query points are drawn from the model distribution $\hat{p}_1$ while database points are drawn from the true target p_1, and we control the resulting density-ratio bias through a Pinsker-plus-Lipschitz argument on the manifold.
- **Empirical scope.** We show that nearest-neighbor replacement alone is insufficient, that partially trained models can already support useful sparse transport plans, and that in weak-to-middle teacher regimes MASS improves retraining efficiency over both independent continuation and drifted model-generated couplings. In stronger teacher regimes, we also identify settings where MASS may help less.

2 Background

The core problem in flow-based generation is to transport a simple source distribution p_0 to a target data distribution p_1. Equivalently, one seeks a transport map or flow map that pushes p_0 onto p_1. This problem is highly non-unique: many maps connect the same marginals. The main question is therefore not only how to find a transport, but how to find one whose trajectories are easy to learn and efficient to integrate at sampling time.

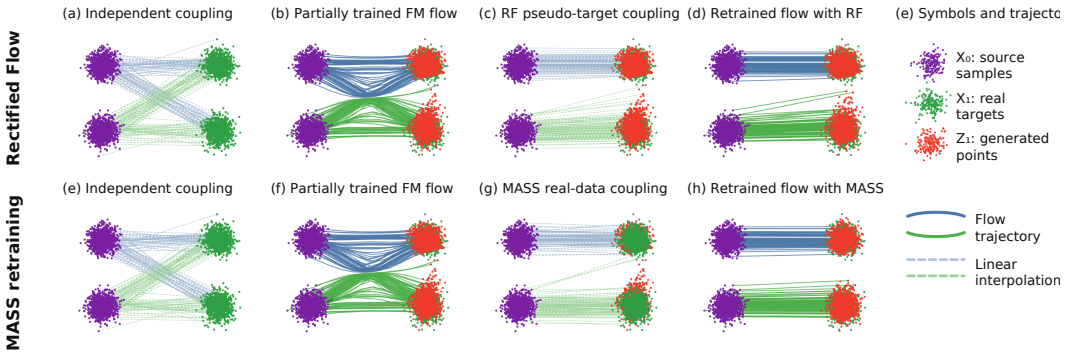

Figure 2: Method schematic contrasting Rectified Flow (top row) and MASS (bottom row) coupling construction from the same warm-start FM model. Panels (a,b,e,f) share the starting point (independent coupling and the FM-trained flow). RF reuses the generated endpoints Z_1 as pseudo-targets (c, d); MASS uses them only as geometric guides and re-couples each source to nearby real targets X_1 (g, h).

2.1 Flow Matching and Transport Maps

Flow Matching (FM) (Lipman et al., 2023) turns the transport problem into regression on a time-dependent vector field $v_\theta(x,t)$. Using a coupling $\pi \in \Pi(p_0,p_1)$ and the linear interpolation $x_t = (1-t)x_0 + tx_1$, Conditional Flow Matching trains

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\,(x_0,x_1)\sim\pi}\big\|v_\theta(x_t,t) - (x_1 - x_0)\big\|^2. \tag{1}$$

FM avoids simulation during training and scales well in modern generative pipelines, but the learned flow depends critically on π. With the standard independent choice $\pi = p_0 \times p_1$ the paired endpoints are often geometrically mismatched, the learned probability path is typically curved, and solving the resulting ODE accurately requires many function evaluations (NFEs).

2.2 Optimal Transport and OT-guided Flow Matching

This non-uniqueness naturally leads to optimal transport (OT), which selects a transport according to a cost. In the static Kantorovich formulation, OT seeks a coupling $\pi \in \Pi(p_0,p_1)$ that minimizes

$$\min_{\pi\in\Pi(p_0,p_1)} \int \tfrac{1}{2}\|x-y\|^2\, d\pi(x,y), \tag{2}$$

in the quadratic case.

OT also admits a dynamic formulation, and under this quadratic cost the optimal map induces straight displacement interpolants between paired points (Villani, 2009). FM itself learns a dynamic transport, but under independent coupling its trajectories are generally not those straight ones, and straighter trajectories would make the learned ODE easier to solve at fewer NFEs — hence the question of how to approximate dynamic OT inside the FM framework.

Tong et al. (Tong et al., 2024) and Pooladian et al. (Pooladian et al., 2023) showed that replacing independent pairs with OT pairs makes FM trajectories straighter in expectation, and proposed minibatch OT, where each training batch solves a discrete OT problem (Figure 3). Quality depends strongly on batch size, while large batches increase memory and computation. Optimal Flow Matching (Kornilov et al., 2024) targets the global quadratic-OT geometry more directly in a single FM step, at the cost of stronger structural constraints and a more demanding training setup. Obtaining straight trajectories robustly and cheaply at dataset scale therefore remains difficult.

2.3 Rectified Flow

Rectified Flow (RF) (Liu et al., 2023) uses the current model itself to define a new coupling: after training FM as a teacher, each source x_0 is paired with its generated endpoint $\hat{x}_1$ and FM is retrained

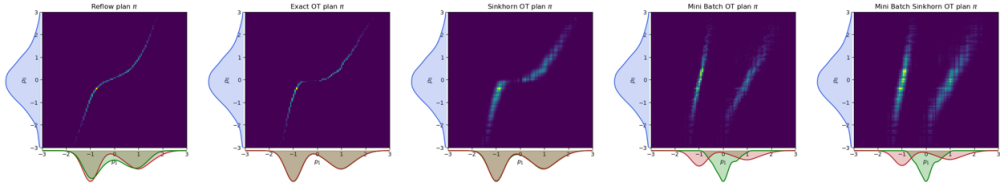

Figure 3: Coupling plans in 1D (barycentric-projection view), left to right: RF coupling, exact OT, Sinkhorn OT, minibatch exact OT, minibatch Sinkhorn OT. The early-stage RF coupling tracks exact OT more closely than minibatch variants, motivating model-aligned sparse OT.

on the model-induced coupling. This straightens trajectories but shifts supervision from real targets to model-generated pseudo-targets, and when the teacher is weak or the ODE solver drifts the endpoints can fall off-manifold and retraining can amplify model error (Figure 2 a–d). Recent RF improvements (Lee et al., 2024; Kim et al., 2025) keep the same intuition that straighter paths help and coupling quality is decisive.

2.4 Mean Flow and its improvements

Mean Flow (MF) (Geng et al., 2025a) learns an average velocity field over a time interval rather than an instantaneous velocity, which makes one-step generation possible. Its training objective $\mathcal{L}_{\mathrm{MF}}(\theta) = \mathbb{E}_{x_0,x_1,r,t}\|u_\theta(x_t,r,t) - \mathrm{sg}(u_{\mathrm{tgt}})\|^2$ uses a target $u_{\mathrm{tgt}} = v(x_t,t) - (t-r)\,\frac{d}{dt}u_\theta(x_t,r,t)$ derived from the Mean Flow identity via a Jacobian-vector product, with $v(x_t,t)$ approximated by the same paired target $x_1 - x_0$ as FM. The JVP correction can be unstable, which has motivated several refinements: improved Mean Flows (iMF) (Geng et al., 2025b) simplify the target and improve conditioning, pixel Mean Flows (pMF) (Lu et al., 2026) extend the idea to one-step latent-free image generation, Decoupled MF (DecMF) (Lee et al., 2026) warm-starts from a pretrained FM, transport-based MF (OTMF) (Akbari et al., 2025) imports OT couplings, and Rectified MF (ReMF) (Zhang et al., 2025) combines MF with rectified couplings from a trained teacher.

Across FM, OT-FM, RF, OTMF, and ReMF the same pattern appears: better couplings produce straighter trajectories, but each method's coupling has a different failure mode. Independent coupling is scalable but curved, OT coupling is principled but expensive, minibatch OT is cheaper but local and strongly batch-size dependent, and rectified coupling can inherit model drift when the teacher is weak. This leaves the central question of our paper: can a partially trained flow already provide enough coarse geometry to construct a scalable, real-data-anchored coupling that better approximates dynamic OT without relying on generated pseudo-targets?

3 MASS: Model-Aligned Sparse Sinkhorn

3.1 When Early Training Becomes Useful

Our central empirical observation is that a flow model can become useful for coupling construction before it becomes a good generator. On CIFAR-10, sample FID drops rapidly from 361.94 to 23.14 in the first 100k iterations, then improves more slowly from 23.14 to 13.42 over the next 100k, and later gains slow further. The 100k checkpoint is therefore not yet a reliable generator, but it is no longer unstructured. Early training quickly moves the model toward the right regions of the target distribution, and later training mainly refines fidelity within those regions.

A 2D moons example shows the same thing (Appendix C.7, Figure 5): nearly uninformative at 10 iterations, the model already places mass near the correct crescents by 500, well before the refinement that arrives at 1000 and $10{,}000$. The useful regime is intermediate — the model points to the right target regions without yet providing reliable pseudo-targets — which raises the question of whether that rough localization signal can be converted into a better coupling.

Algorithm 1 MASS plan construction

Input: warm-up model $\hat{f}_\theta$, source $\{x_{0,i}\}_{i=1}^N$, dataset $\mathcal{D}=\{x_{1,j}\}_{j=1}^N$, neighborhood k, η, τ, NFEs m
for $i = 1$ **to** N **do**
  $\hat{x}_{1,i} \leftarrow \hat{f}_\theta(x_{0,i})$ [m NFE steps]
  $\mathcal{N}_i^k \leftarrow \mathrm{kNN}(\hat{x}_{1,i}, \mathcal{D}, k)$
end for
$G_k \leftarrow \{(i,j) : j \in \mathcal{N}_i^k\}$
$C_{ij} \leftarrow \|x_{0,i} - x_{1,j}\|^2$ for $(i,j) \in G_k$
$\tilde{\Pi}^k \leftarrow \text{Log-SR-Sinkhorn}(C, \frac{1}{N}\mathbf{1}, \frac{1}{N}\mathbf{1}, \eta, \tau)$
Row-normalize each nonempty row of $\tilde{\Pi}^k$
Return: $\tilde{\Pi}^k$

Algorithm 2 Stochastic sampling from $\tilde{\Pi}^k$

Input: row-stochastic plan $\tilde{\Pi}^k$, dataset $\mathcal{D}$
for $i = 1$ **to** N **do**
  $p_i \leftarrow$ nonzero entries of row i
  $J_i \sim \mathrm{Categorical}(p_i)$
  $\tilde{x}_{1,i} \leftarrow x_{1,J_i}$
end for
Return: $\{(x_{0,i}, \tilde{x}_{1,i})\}_{i=1}^N$

3.2 Why Nearest-Neighbor Correction Is Not Enough

The most immediate correction is nearest-neighbor replacement: give each generated endpoint $\hat{x}_{1,i}$ its nearest training sample as the corrected target, so that every target is real data rather than a model-generated pseudo-target. This solves the wrong problem. It improves pointwise realism while ignoring the induced target distribution: some datapoints receive many assignments and others none, so the corrected target distribution can drift even though every assigned point is real. The failure is clearest for the strong RF 800k teacher of Table 1, where nearest-neighbor correction worsens the corrected-target FID from 2.74 to 14.85; on moons it improves W_2 from 0.411 to 0.374 and still trails the later 1000-iteration model.

What is needed instead is a method that keeps the local proposal made by the early-stage model but reweights mass globally, so that the corrected target distribution drifts as little as possible. Rather than committing to one nearest neighbor, MASS keeps a larger model-guided neighborhood and solves for weights that preserve the target distribution. This changes the outcome sharply: CIFAR-10 corrected-target FID from 23.14 to 0.61, and 2D transport error from about 0.41 to about 0.18.

3.3 Algorithm

MASS turns early-stage model outputs into a reusable sparse transport plan. The model provides a neighborhood proposal, and OT turns that proposal into a sparse real-data coupling. The neighborhood size k controls how narrowly or broadly MASS trusts the proposal.

Algorithm 1 summarizes the four steps. We first generate N approximate model samples $\hat{x}_{1,i} = \hat{f}_\theta(x_{0,i})$ using m NFEs. For each $\hat{x}_{1,i}$ we retrieve its k nearest neighbors $\mathcal{N}_i^k$ in the training dataset $\mathcal{D} = \{x_{1,j}\}_{j=1}^N$, and define the bipartite graph $G_k = \{(i,j) : j \in \mathcal{N}_i^k\}$. The cost matrix is $C_{ij} = \|x_{0,i} - x_{1,j}\|^2$ on G_k and $+\infty$ elsewhere; a deterministic nearest-neighbor rule commits each source to a single target and generally distorts the target marginal, whereas MASS preserves a set of plausible matches and lets OT determine their weights. Because the sparse graph may leave some target points uncovered, the standard Sinkhorn solver can be infeasible, and we instead apply the Semi-Relaxed Robust Sinkhorn algorithm (Le et al., 2021) (Appendix B), which fixes the source marginal and penalizes the target marginal's deviation by a KL term with robustness parameter τ. The output is a sparse plan $\tilde{\Pi}^k$ supported on G_k.

During training, we draw one target per source from the fixed row-normalized plan once per epoch using a Gumbel-max sampler. This decouples the deterministic plan, which fixes the global transport geometry once and for all, from the stochastic per-epoch realization, which supplies the variation in (x_0, x_1) pairs that flow training requires. Relative to a single hard matching it adds regularization analogous to data augmentation while preserving marginal consistency, and it retains the geometric structure that is lost under fully independent coupling. The neighborhood size k acts as a coupling knob; small k yields a sharper but riskier plan, while larger k improves coverage and moves MASS closer to global OT at higher computational cost. The next section addresses how large k must be for the plan to provably match an external quality baseline.

3.4 MARGINAL QUALITY ANALYSIS

Let $\hat{p}_1 = \hat{T}_{\#}p_0$ be the pretrained model's output distribution and p_1 the true target. For numerical endpoint generation, $\hat{T}$ denotes the finite-m solver map actually used to build the graph. Its output distribution therefore includes both model error and integration error; the density assumptions below must hold for this distribution. We fix $\Delta_{\mathrm{KL}} = D_{\mathrm{KL}}(\hat{p}_1 \| p_1)$ and $\Delta_W = W_2(\hat{p}_1, p_1)$ as baseline quality levels for the pretrained model. The question is whether the MASS target marginal $\mathbf{b}^k$, which lives on the empirical target cloud $\{Y_j\}$, can be certified to fall below these baselines.

The key quantity is the *graph defect*

$$\epsilon_k = \min_{T \in \mathcal{F}_k} D_{\mathrm{KL}}(T^\top \mathbf{1}_N \| \mathbf{b}),$$

the minimum KL any plan supported on the k-NN graph can achieve against the uniform target marginal $\mathbf{b}$. It satisfies $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_N = 0$, so more neighbors always help, and at $k = N$ the full bipartite graph admits a plan with exact marginals. Complete proofs are in Appendix A.

Theorem 3.1 (informal; Theorem A.5)**.** *If $\epsilon_k < \Delta_{\mathrm{KL}}$, then for sufficiently large τ, $D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b}) < \Delta_{\mathrm{KL}}$.*

Theorem 3.2 (informal; Theorem A.7)**.** *If $\epsilon_k < 2(\Delta_W / D)^4$, where D is the target-cloud diameter, then for sufficiently large τ, $W_2(\bar{p}_1^k, \bar{p}_1) < \Delta_W$.*

These two results reduce the question of marginal quality to controlling ϵ_k. The following theorem makes this actionable by providing explicit thresholds on k.

Theorem 3.3 (informal; Theorem A.12)**.** *Under Assumption A.1, let $\Delta_\star = \Delta_{\mathrm{KL}} - a_3\, \Delta_{\mathrm{KL}}^{1/(d+1)} > 0$, $\varepsilon \in (0,1)$ and define*

$$k_0 = \max\left\{ \left\lceil C_0 \log \tfrac{N}{\varepsilon} \right\rceil,\ \left\lceil \frac{2a_1 \log(N/\varepsilon)}{\Delta_\star} \right\rceil \right\},\quad k_{\max} = \min\left\{ \lfloor \gamma N \rfloor,\ \left\lfloor \left(\frac{\Delta_\star}{2a_2} \right)^{d/2} N \right\rfloor \right\},$$

where $a_1, a_2, a_3, C_0, \gamma$ are constants depending on manifold geometry and density bounds. If $k_0 \leq k_{\max}$, then with probability at least $1 - \varepsilon$, every $k \geq k_0$ satisfies $\epsilon_k < \Delta_{\mathrm{KL}}$.

Remark 3.4 (Scope and computability). The constants depend on intrinsic-dimension, density, and curvature quantities not accessible in practice, and the requirement $\Delta_\star > 0$ tightens as d grows, so we read Theorem 3.3 as a low-dimensional certificate and a structural guide for choosing k rather than a tight image-scale bound, and select k by a proxy quality metric (Section 4.1). Remark A.11 gives the full discussion; Appendix A.6 gives a setting in which every constant is explicit and the graph defect reaches $\epsilon_k = 0$ at a threshold we can measure.

4 EXPERIMENTS

We focus the main text on CIFAR-10 (Krizhevsky, 2009), where our two main empirical questions are whether MASS improves the fixed coupling before retraining and whether that improvement carries over to continuation from a partially trained checkpoint. Full CIFAR-10 training, MASS construction, and evaluation details are given in Appendix C.1, and additional 2D toy experiments in Appendix C.7; further ablations and diagnostics are in Appendix C.

4.1 UNDERSTANDING MASS BEFORE RETRAINING

Before studying retraining, we isolate the fixed coupling on CIFAR-10 and ask whether the warm-up model already provides enough coarse geometry to define useful candidate neighborhoods. Starting from FM checkpoints at step 0, 50k, and 100k, we sample targets directly from the frozen MASS plan (Algorithm 2) and measure the induced target-distribution FID. We also evaluate an 800k-step Rectified Flow checkpoint from the original repository as a strong reference. The experiment addresses three questions. Does the warm-up model become more useful as training proceeds? Is naive nearest-neighbor replacement enough? How much does increasing k help?

Table 1 answers the first question positively. As warm-up training proceeds from FM 0 to FM 50k to FM 100k, the same architecture becomes a much better coupling constructor; at fixed $k = 200$ the

Table 1: CIFAR-10 fixed-coupling quality before retraining (lower is better). "Teacher" is the FM checkpoint's own sample FID; "NN" is naive k=1 nearest-neighbor replacement; MASS columns sample directly from the frozen sparse plan. The bottom row reports the metric's intrinsic floor (uniform with-replacement subsample of train, N=50,000, mean $\pm$ std over 5 seeds).

Checkpoint	Teacher	NN	MASS k=50	MASS k=100	MASS k=200	MASS k=500
FM 0	450.92	339.45	115.40	86.42	62.29	39.30
FM 50k	117.99	74.13	17.91	14.13	10.76	6.81
FM 100k	23.14	21.75	1.94	1.03	0.62	0.50
RF 800k	2.74	14.85	1.03	0.76	0.66	0.61
Reference floor (uniform with-replacement subsample of train, N=50,000): FID $= 0.528 \pm 0.003$						

induced-target FID improves from 62.29 to 10.76 to 0.62. This is the main CIFAR-10 evidence that a partially trained model can already be useful before it generates well. The same row also makes a stronger point about robustness to teacher quality. Even at FM 0, where the teacher is essentially untrained with sample FID 450.92, MASS at $k = 500$ already reaches FID 39.30, lower than the FM-50k teacher's own sample quality. The kNN backbone with reweighting therefore extracts useful geometric signal even from a near-random teacher, so MASS does not require a strong teacher to produce a usable coupling.

Nearest-neighbor replacement can already improve over the teacher in the FM rows, so simple snapping is not useless. MASS is nevertheless consistently much stronger. At FM 100k, nearest-neighbor replacement gives FID 21.75, whereas MASS reaches 1.94 at $k = 50$, 1.03 at $k = 100$, and 0.62 at $k = 200$. The same conclusion holds for the strong RF 800k teacher, where nearest-neighbor replacement gives 14.85 while MASS remains below 1.1 across the tested k values. The gain comes from using a broader candidate set and reweighting it globally, not from nearest-neighbor snapping alone.

Larger neighborhoods consistently improve the transported distribution, with diminishing returns once the teacher is informative. At FM 100k, FID drops from 1.94 at $k = 50$ to 1.03 at $k = 100$ and 0.62 at $k = 200$, but only to 0.50 at $k = 500$; the same pattern appears for the strong RF teacher ($1.03 \to 0.76 \to 0.66 \to 0.61$). This supports k as a locality versus coverage knob and suggests that on CIFAR-10 most of the benefit is already captured by $k \approx 100$ to 200.

Together, the table establishes the claim needed before the training curves. In the controlled CIFAR-10 setting, partially trained models become progressively more useful for coupling construction, nearest-neighbor correction alone is insufficient, and MASS improves when given a broader candidate set. Whether better fixed coupling also yields faster training is tested next.

The FID in Table 1 cannot reach zero. Even a target set drawn uniformly at random from the training images themselves — the best any plan could do — scores FID $= 0.528 \pm 0.003$ at N=50,000, because FID is estimated from finitely many samples (Appendix C.1). Two consequences follow. On the FM-100k and RF-800k rows, MASS at $k \geq 200$ already sits at that floor, so the remaining difference between k=200 and k=500 is not informative; both plans recover near-uniform sampling of the dataset. The distance between NN (21.75, 14.85) and MASS (roughly 0.5 to 1) on the same rows is not a floor artifact, however, but a real $\sim 30\times$ effect, caused by NN's marginal collapse — the failure mode the threshold-k certificate is designed to control.

A target set and a coupling are not the same thing, and Table 1 scores only the first. Appendix C.2 scores four alternative constructions both ways and separates the contributions of marginal correction, real-data anchoring, and model alignment.

4.2 Retraining with MASS

We study whether the coupling from Section 4.1 also improves retraining. All runs start from the same 100k-step FM checkpoint. For Figure 4(a), we compare direct FM continuation against two warm-started variants. We first generate 50k samples from the 100k checkpoint with 10 NFE. ReFM then continues training with generated-data coupling, while MASS ReFM uses a MASS plan built from the same checkpoint with $k = 200$. For Figure 4(b), we switch to iMF retraining from the

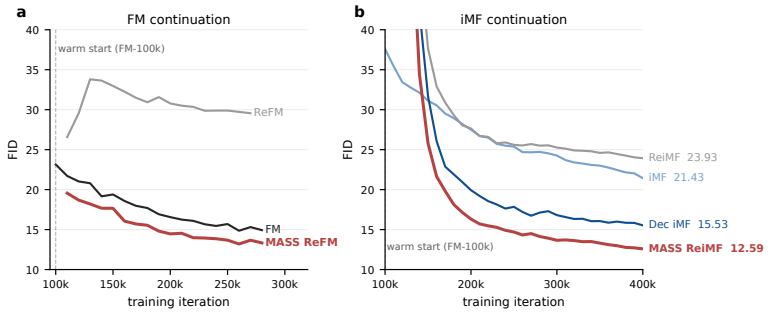

Figure 4: CIFAR-10 retraining curves from a shared FM-100k warm start. (a) FM continuation with RF-style and MASS-based coupling. (b) iMF-based continuation comparing ReiMF, Dec-iMF, and MASS+ReiMF; FID values at 400k iterations are shown next to the endpoint labels.

same warm start, since iMF is more stable than the original MF objective. Dec-iMF continues with independent coupling, ReiMF uses the generated samples, and MASS ReiMF uses the MASS plan.

Figure 4(a) shows that generated-data coupling is not reliable enough for FM continuation from this checkpoint. ReFM drifts away, whereas MASS ReFM remains slightly but consistently better than direct FM continuation. Figure 4(b) shows the same pattern more clearly under iMF retraining. At 400k total iterations, iMF reaches FID 21.43, Dec-iMF reaches 15.53, ReiMF reaches about 23.92, and MASS ReiMF reaches 12.59. The main effect therefore comes from the coupling. Reusing the same warm start with a MASS plan is more effective than either independent coupling or generated-data retraining.

Strong-teacher regime. Rerunning the comparison from an RF-800k teacher, 1-NFE ReFM reaches FID 11.62 after 100k iterations while MASS ReFM only catches up at 4 NFE (13.49): the near-straight teacher already produces good pseudo-pairs and the MASS plan adds variance, so the benefit is confined to the partially trained regime.

4.3 Single-cell trajectory interpolation

Nothing in the MASS construction is specific to images, so we also test it where the transport carries a physical meaning. The embryoid-body dataset records single-cell gene expression at five developmental timepoints; the task is to learn the flow between consecutive timepoints and then predict the distribution of cells at a timepoint the model never saw. We follow the preprocessing, leave-one-timepoint-out protocol and evaluation of Tong et al. (2024), holding out timepoints 1, 2, and 3 in turn and scoring the predicted distribution against the held-out cells by 1-Wasserstein distance. Only the coupling differs between runs: all four share a 2k-step conditional-FM warm-up under independent pairing, then continue for 5k steps with their own coupling. As a check on our implementation of the protocol, our minibatch-OT run reproduces the value reported by Tong et al. (2024) (0.785 ± 0.007 against their 0.790 ± 0.068). Over five seeds MASS reaches 0.764 ± 0.010, ahead of independent pairing (0.900 ± 0.032) and ahead of minibatch OT in every seed, and level with exact global OT (0.768 ± 0.019), which is affordable here only because this dataset is small enough to solve exactly. The conclusion is a bounded one: an early-stage model can build a useful coupling outside image generation, and on this benchmark a plan built once reaches the quality of exact OT rather than improving on it.

5 Related Work

MASS sits between three lines of work. **OT-guided Flow Matching** replaces independent pairings with transport-aware ones to straighten trajectories (Tong et al., 2024; Pooladian et al., 2023; Kornilov et al., 2024; Akbari et al., 2025), but solves a fresh OT problem per minibatch and gives no fixed dataset-scale coupling for continuation training. **Rectified and Mean Flow** use the model itself to define pairings: Rectified Flow retrains on teacher-generated pairs (Liu et al., 2023; Lee et al., 2024; Kim et al., 2025), and Mean Flow variants refine the objective (Zhang et al., 2025), but both inherit teacher drift when the teacher is weak. Closest to MASS is **Model-Aligned Coupling (MAC)** (Lin

et al., 2025), which scores candidate matches with the current model *during* training; we instead invoke a partially trained model *once* to fix a sparse candidate graph, so retraining costs match standard FM. A third line is **Sparse OT**: semi-relaxed Sinkhorn (Le et al., 2021; Fukunaga & Kasai, 2022) relaxes one marginal but keeps a dense cost, and Spar-Sink (Li et al., 2023) accelerates entropic OT by *randomly* sparsifying the Sinkhorn kernel, which shifts both marginals and the resulting plan, and was not proposed for flow training. MASS instead uses *structural*, model-guided sparsity: the partially trained flow decides which pairs are eligible at all, and semi-relaxed Sinkhorn keeps the plan anchored to the real source marginal.

Table 2: Where MASS sits among coupling-construction methods for flow training. "Once" is a one-time offline construction reused throughout training; "per batch" is a fresh transport problem at every step. N is dataset size, B is batch size, k is the MASS neighborhood size.

Method	Targets	When constructed	Cost	Geometric scope
Independent FM (Lipman et al., 2023)	real	–	$O(1)$	none
Rectified Flow (Liu et al., 2023)	pseudo	once	$O(N \cdot \mathrm{highNFE})$	global (model)
Minibatch OT-FM (Tong et al., 2024)	real	per batch	$O(B^2)$ per step	local (minibatch)
Optimal FM (Kornilov et al., 2024)	real	per step	high (parametric)	global
MAC (Lin et al., 2025)	real	per batch	$O(B^2)$ per step	local (model-scored)
MASS (ours)	real	once	$O(N \cdot \mathrm{lowNFE} + kN)$	global (model-guided)

The new combination in MASS is to build a *one-shot*, real-data-anchored sparse OT plan from an *early-stage* teacher's structural hints rather than from a fully trained generator, per-batch model scoring, or random kernel sparsification, and to certify when the resulting target marginal improves over the teacher's output distribution via a two-measure k-NN degree concentration argument.

6 Conclusion

MASS uses a partially trained flow model only as a geometric guide, and reanchors coupling construction to real data through semi-relaxed Sinkhorn reweighting on a sparse k-NN graph. The plan is built once, costs $O(kN)$, and composes with any flow-matching objective. On CIFAR-10, MASS substantially improves coupling quality before retraining, turning a partially trained checkpoint into a usable coupling constructor, and it reduces continuation cost relative to both independent FM and Rectified-Flow-style model-generated coupling, especially in the weak-to-moderate teacher regime. Once the teacher is already strong, however, its pseudo-pairs are nearly straight and the MASS plan only adds variance, so MASS is best read as a tool for the partially trained regime rather than a drop-in replacement for late-stage retraining of an already strong flow.

The threshold-k certificate (Theorem 3.3) is a low-dimensional guarantee, since the $\Delta_{\mathrm{KL}}^{1/(d+1)}$ term degrades with intrinsic dimension, and we therefore read the theory as a structural guide for k-selection rather than a tight image-scale bound. Extending MASS to higher-resolution datasets, latent flows, and class-conditional generation are natural next steps.

AI use statement

The research idea, the theoretical framework, the proofs, and the design of every experiment reported here are the authors' own. Within that scope we used generative AI assistants (Anthropic's Claude, OpenAI Codex, and ChatGPT) for the following tasks that ICLR requires us to disclose. It helped sharpen the statements of the theorems and their assumptions, checked our written proof steps for gaps, and edited the proof write-up, but supplied no lemma, bound, or proof step on which the results depend. It debugged our code and wrote some auxiliary pipeline, evaluation, and plotting components, all of which we read and ran against our own reference implementation before producing any reported number with them. It drafted some of the wording in which the measurements are interpreted, which we checked against the underlying numbers. And it helped render into English parts of the text. We did not use it to generate synthetic data sets, to develop the conceptual framework, to propose or refine hypotheses, to design the methodology or the experiments or to give feedback on them, or to clean or reformat datasets. In the recommended categories we also used it for code editing, for copy-editing. Every number, table, and figure here, including the teaser, is the output of an actual run of our code; no result, image, or data point was generated by a model. We have reviewed all

AI-assisted work and take responsibility for the final content of this submission, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The MASS construction is given as pseudocode in Algorithms 1 and 2 and the solver in Appendix B; Appendix A states every assumption and gives complete proofs of the claims. Appendix C gives the training configuration, plan hyperparameters, data processing, and evaluation protocol for every experiment reported here. Source code and the MASS construction scripts are provided in the supplementary material; we plan to release the pretrained checkpoints in a public repository upon acceptance.

References

Elaheh Akbari, Shansita Sharma, Ping He, Ahmadreza Moradipari, Kyungtae Han, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. OT-MeanFlow3D: Bridging optimal transport and Meanflow for efficient 3D point cloud generation. *arXiv preprint arXiv:2509.22592*, 2025. URL https://arxiv.org/abs/2509.22592.

Takumi Fukunaga and Hiroyuki Kasai. On the convergence of semi-relaxed sinkhorn with marginal constraint and ot distance gaps. *arXiv preprint arXiv:2205.13846*, 2022. URL https://arxiv.org/abs/2205.13846.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6d13e085b79d454da5910e4ca82a3d9d-Abstract-Conference.html.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b. URL https://arxiv.org/abs/2512.02012.

Shin-seong Kim, Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Balanced conic rectified flow. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kqHzgTV9AU.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc8f76d9caadd48f77025b1c889d2e2d-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Khang Le, Huy Nguyen, Quang M. Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b80ba73857eed2a36dc7640e2310055a-Abstract.html.

Kyungmin Lee, Sihyun Yu, and Jinwoo Shin. Decoupled MeanFlow: Turning flow models into flow maps for accelerated sampling. *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=P4VsQexEPe.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7343a5c976f8399880b695267f1f9e9f-Abstract-Conference.html.

Mengyu Li, Jun Yu, Tao Li, and Cheng Meng. Importance sparsification for Sinkhorn algorithm. *Journal of Machine Learning Research*, 24(247):1–44, 2023. URL https://www.jmlr.org/papers/v24/22-1311.html.

Yexiong Lin, Yu Yao, and Tongliang Liu. Beyond optimal transport: Model-aligned coupling for flow matching. *arXiv preprint arXiv:2505.23346*, 2025. URL https://arxiv.org/abs/2505.23346.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026. URL https://arxiv.org/abs/2601.22158.

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 *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

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. URL https://openreview.net/forum?id=CD9Snc73AW.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. URL https://doi.org/10.1007/978-3-540-71050-9.

Xinxi Zhang, Shiwei Tan, Quang Nguyen, Quan Dao, Ligong Han, Xiaoxiao He, Tunyu Zhang, Chengzhi Mao, Dimitris Metaxas, and Vladimir Pavlovic. Overcoming the curvature bottleneck in MeanFlow. *arXiv preprint arXiv:2511.23342*, 2025. URL https://arxiv.org/abs/2511.23342.

A THEORETICAL ANALYSIS OF THE THRESHOLD-k CERTIFICATE

This appendix gives complete statements and proofs for the MASS threshold-k certificate: a sufficient number of nearest-neighbor candidates k at which the MASS empirical transport certificate falls below externally supplied baseline quality levels for the pretrained model, in both KL divergence and W_2 distance. The KL statements below are discrete statements about the MASS target marginal on the observed target cloud; they do not require estimating a continuous KL divergence from two empirical point clouds.

A.1 SETUP AND NOTATION

Let $(\mathcal{M}, g)$ be a compact d-dimensional Riemannian manifold without boundary, embedded in $\mathbb{R}^D$. Let $\mathrm{vol}_{\mathcal{M}}$ denote the Riemannian volume measure on $\mathcal{M}$. Let p_0 be a probability measure on $\mathbb{R}^D$, and let $p_1, \hat{p}_1$ be probability measures supported on $\mathcal{M}$, with densities f and $\hat{f}$ with respect to $\mathrm{vol}_{\mathcal{M}}$.

Let $\hat{T} : \mathbb{R}^D \to \mathcal{M}$ be a pretrained flow model and $\hat{p}_1 = \hat{T}_{\#}p_0$. Given i.i.d. source samples $X_1, \dots, X_N \sim p_0$, define the *model-generated samples*

$$\hat{Y}_i = \hat{T}(X_i), \qquad i = 1, \dots, N.$$

Let $Y_1, \dots, Y_N \sim p_1$ be i.i.d. target samples on $\mathcal{M}$, independent of $\{X_i\}$. The corresponding empirical measures are

$$\bar{p}_0 = \tfrac{1}{N}\sum_{i=1}^N \delta_{X_i}, \qquad \bar{p}_1 = \tfrac{1}{N}\sum_{j=1}^N \delta_{Y_j},$$

with uniform weight vectors $\mathbf{a} = \mathbf{b} = \frac{1}{N}\mathbf{1}_N$. Distances between points on $\mathcal{M}$ are measured in the ambient Euclidean norm $\|\cdot\|$ of $\mathbb{R}^D$.

Graph and sparse plan. For $k \in \{1, \dots, N\}$ define the k-nearest-neighbor index sets and bipartite graph

$$\mathcal{N}_i^k = \mathrm{kNN}(\hat{Y}_i; \{Y_1, \dots, Y_N\}; k), \qquad G_k = \{(i,j) : j \in \mathcal{N}_i^k\}. \tag{3}$$

By construction $G_1 \subseteq G_2 \subseteq \cdots \subseteq G_N$, where G_N is the complete bipartite graph. Note $|G_k| \le Nk$.

The feasible transport sets are

$$\mathcal{F} = \{T \ge 0 : T\mathbf{1}_N = \mathbf{a}\}, \qquad \mathcal{F}_k = \{T \in \mathcal{F} : T_{ij} = 0,\ \forall\, (i,j) \notin G_k\}.$$

Cost matrix. The transport cost between source sample X_i and target Y_j is

$$C_{ij} = \|X_i - Y_j\|^2, \qquad (i,j) \in G_k, \tag{4}$$

and $C_{ij} = +\infty$ for $(i,j) \notin G_k$. Denote the maximum restricted cost by

$$\lambda_k = \max_{(i,j)\in G_k} \|X_i - Y_j\|^2. \tag{5}$$

MASS objective. The MASS semi-relaxed entropic OT problem is

$$T^k = \arg\min_{T \in \mathcal{F}_k} \Big\{ \langle C, T\rangle + \tau\, D_{\mathrm{KL}}(T^\top \mathbf{1}_N \| \mathbf{b}) - \eta H(T) \Big\}, \tag{6}$$

where $H(T) = -\sum_{ij} T_{ij} \log T_{ij} \ge 0$ is the Shannon entropy and $\tau, \eta > 0$ are regularization parameters. The solution T^k is unique by strict convexity. We denote its target marginal and the induced transported measure by

$$\mathbf{b}^k = (T^k)^\top \mathbf{1}_N, \qquad \tilde{p}_1^k = \textstyle\sum_j b_j^k\, \delta_{Y_j}. \tag{7}$$

Uniform reference plan. Define the uniform plan supported on G_k:

$$U_{ij}^k = \tfrac{1}{Nk}\, \mathbf{1}\{j \in \mathcal{N}_i^k\}, \tag{8}$$

with target marginal $q_j = d_j(k)/(Nk)$, where $d_j(k) = |\{i : j \in \mathcal{N}_i^k\}|$ is the in-degree of Y_j in G_k.

Graph defect.

$$\epsilon_k = \min_{T \in \mathcal{F}_k} D_{\mathrm{KL}}(T^\top \mathbf{1}_N \| \mathbf{b}). \tag{9}$$

Since $U^k \in \mathcal{F}_k$ with marginal q, we have $\epsilon_k \le D_{\mathrm{KL}}(q\|\mathbf{b})$ immediately.

Baseline quality levels. Let $\Delta_{\mathrm{KL}} > 0$ and $\Delta_W > 0$ be externally supplied baseline quality levels for the pretrained model. They may be population discrepancies, validated estimates, or conservative tolerances against which the empirical MASS certificate is compared. In the ideal population notation one may think of

$$\Delta_{\mathrm{KL}} = D_{\mathrm{KL}}(\hat{p}_1 \| p_1), \qquad \Delta_W = W_2(\hat{p}_1, p_1). \tag{10}$$

The results use only the scalar values Δ_{KL} and Δ_W; they do not require evaluating $D_{\mathrm{KL}}(\hat{p}_1\|p_1)$ from the finite point clouds.

Target cloud diameter.

$$D = \max_{j,l \in [N]} \|Y_j - Y_l\|. \tag{11}$$

Assume $D > 0$ and set $\rho = \Delta_W / D$.

Assumption A.1. The densities f and $\hat{f}$ with respect to $\mathrm{vol}_{\mathcal{M}}$ satisfy

$$0 < m \le f(y),\ \hat{f}(y) \le M < \infty \quad \text{for all } y \in \mathcal{M},$$

and are Lipschitz in the Riemannian distance $d_{\mathcal{M}}$:

$$|f(x) - f(y)| \le L\, d_{\mathcal{M}}(x,y) \quad \text{and} \quad |\hat{f}(x) - \hat{f}(y)| \le L\, d_{\mathcal{M}}(x,y) \quad \text{for all } x, y \in \mathcal{M}.$$

A.2 Graph Defect: KL Bounds

Lemma A.2. $\epsilon_1 \ge \epsilon_2 \ge \cdots \ge \epsilon_N = 0$.

Proof. Since $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$, the infimum over the larger set is no larger: $\epsilon_{k+1} \le \epsilon_k$. At $k = N$: the uniform plan $T_{ij} = 1/N^2$ belongs to $\mathcal{F}_N$, has target marginal $\mathbf{b}$, and gives $D_{\mathrm{KL}}(\mathbf{b}\|\mathbf{b}) = 0$, so $\epsilon_N = 0$. □

Lemma A.3. *For any $T \ge 0$ with $\sum_{ij} T_{ij} = 1$ and $\mathrm{supp}(T) \subseteq G_k$,*

$$0 \le H(T) \le \log(Nk).$$

Proof. Non-negativity holds since $T_{ij} \log T_{ij} \le 0$ for $T_{ij} \in (0,1)$. Among distributions on at most $|G_k| \le Nk$ atoms, entropy is maximized by the uniform distribution, giving $H(T) \le \log(Nk)$. □

Theorem A.4.

$$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) \ \le\ \epsilon_k + \frac{\lambda_k + \eta \log(Nk)}{\tau}.$$

Proof. Let $T_* \in \mathcal{F}_k$ achieve $D_{\mathrm{KL}}(T_*^\top \mathbf{1}\|\mathbf{b}) = \epsilon_k$. Since T^k minimizes equation 6 over $\mathcal{F}_k$:

$$\langle C, T^k\rangle + \tau\, D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) - \eta H(T^k) \ \le\ \langle C, T_*\rangle + \tau\,\epsilon_k - \eta H(T_*).$$

Since $C_{ij} \le \lambda_k$ on G_k and the entries of T_* sum to 1, we have $\langle C, T_*\rangle \le \lambda_k$. Using $\langle C, T^k\rangle \ge 0$, $H(T^k) \le \log(Nk)$, and $H(T_*) \ge 0$:

$$\tau\, D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) \ \le\ \tau\,\epsilon_k + \lambda_k + \eta \log(Nk).$$

Dividing by $\tau > 0$ gives the result. □

Theorem A.5. *Suppose $\epsilon_k < \Delta_{\mathrm{KL}}$. Then for any*

$$\tau \ >\ \frac{\lambda_k + \eta \log(Nk)}{\Delta_{\mathrm{KL}} - \epsilon_k},$$

we have $D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) < \Delta_{\mathrm{KL}}$.

Proof. By Theorem A.4:

$$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) \le \epsilon_k + \frac{\lambda_k + \eta\log(Nk)}{\tau} < \epsilon_k + (\Delta_{\mathrm{KL}} - \epsilon_k) = \Delta_{\mathrm{KL}}.$$ □

A.3 W_2 Bounds via Graph Defect

Corollary A.6.

$$W_2(\tilde{p}_1^k, \bar{p}_1) \;\le\; D\left(\frac{1}{2}\left(\epsilon_k + \frac{\lambda_k + \eta\log(Nk)}{\tau}\right)\right)^{1/4}.$$

Proof. Construct a coupling by matching mass $\min(b_j^k, b_j)$ at each Y_j at zero cost, and transporting the residual mass within the target cloud. Since all atoms lie in $\{Y_1, \dots, Y_N\}$, the squared cost of any residual pair is at most D^2:

$$W_2^2(\tilde{p}_1^k, \bar{p}_1) \;\le\; D^2 \cdot \mathrm{TV}(\tilde{p}_1^k, \bar{p}_1). \tag{12}$$

Pinsker's inequality gives $\mathrm{TV}(\tilde{p}_1^k, \bar{p}_1) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b})}$. Substituting into equation 12, applying Theorem A.4, and taking square roots gives the result. □

Theorem A.7. *Suppose $\epsilon_k < 2\rho^4$. Then for any*

$$\tau \;>\; \frac{\lambda_k + \eta\log(Nk)}{2\rho^4 - \epsilon_k},$$

we have $W_2(\tilde{p}_1^k, \bar{p}_1) < \Delta_W$.

Proof. By Corollary A.6:

$$W_2(\tilde{p}_1^k, \bar{p}_1) \le D\left(\frac{1}{2}\left(\epsilon_k + \frac{\lambda_k + \eta\log(Nk)}{\tau}\right)\right)^{1/4} < D\left(\frac{\epsilon_k + (2\rho^4 - \epsilon_k)}{2}\right)^{1/4} = D\rho = \Delta_W.$$

□

A.4 Degree Concentration and Defect Control

We now bound the graph defect ϵ_k probabilistically. In the MASS graph, the query points $\hat{Y}_i \sim \hat{p}_1$ and the database points $Y_j \sim p_1$ are drawn from *different* distributions. The in-degree $d_j(k) = |\{i : j \in \mathcal{N}_i^k\}|$ counts how often Y_j is selected as a nearest neighbor of some $\hat{Y}_i$. When $\hat{p}_1 \neq p_1$, the ratio $d_j(k)/k$ concentrates around $\hat{f}(Y_j)/f(Y_j)$ rather than 1; controlling this gap requires a two-measure concentration argument on $\mathcal{M}$ together with a density-ratio bound derived from Δ_{KL}.

Lemma A.8.

$$\epsilon_k \;\le\; D_{\mathrm{KL}}(q\|\mathbf{b}) \;\le\; \frac{1}{N}\sum_{j=1}^{N}\left(\frac{d_j(k)}{k} - 1\right)^2.$$

Proof. The uniform plan U^k belongs to $\mathcal{F}_k$ with target marginal q, so $\epsilon_k \le D_{\mathrm{KL}}(q\|\mathbf{b})$ by definition of the infimum.

For the second inequality, set $x_j = d_j(k)/k$. Since each row of G_k has exactly k entries, $\sum_j d_j(k) = Nk$, hence $\frac{1}{N}\sum_j x_j = 1$. Using $\log x \le x - 1$ for $x > 0$:

$$x_j \log x_j \le x_j(x_j - 1) = x_j^2 - x_j.$$

Therefore

$$\begin{aligned} D_{\mathrm{KL}}(q\|\mathbf{b}) &= \frac{1}{N}\sum_j x_j \log x_j \le \frac{1}{N}\sum_j (x_j^2 - x_j) \\ &= \frac{1}{N}\sum_j (x_j - 1)^2 + \underbrace{\frac{2}{N}\sum_j (x_j - 1)}_{=0} = \frac{1}{N}\sum_j \left(\frac{d_j(k)}{k} - 1\right)^2. \end{aligned}$$

□

Let $\omega_d = \pi^{d/2}/\Gamma(d/2+1)$ denote the volume of the unit ball in $\mathbb{R}^d$. On the compact manifold $\mathcal{M}$, the volume of a geodesic ball of radius r satisfies $\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(x,r)) = \omega_d r^d\big(1+O(r^2)\big)$ uniformly in x, with the $O(r^2)$ term controlled by the Ricci curvature of $\mathcal{M}$.

Theorem A.9. *Under Assumption A.1, for any $\delta \in (0,1)$ and any k such that $C_0 \log(N/\delta) < k < \gamma N$, with probability at least $1-\delta$:*

$$\max_{1\le j\le N}\left|\frac{d_j(k)}{k}-1\right| \;\le\; C_1\sqrt{\frac{\log(N/\delta)}{k}} + C_2\left(\frac{k}{N}\right)^{1/d} + C_3\,\Delta_{\mathrm{KL}}^{\frac{1}{2(d+1)}}, \tag{13}$$

where $\gamma, C_0, C_1, C_2, C_3 > 0$ are constants depending only on m, M, L, d, ω_d, and the geometry of $\mathcal{M}$.

Proof. Set $L_N = \log(N/\delta)$. Throughout, constants may change from line to line but depend only on m, M, L, d, ω_d, and the geometry of $\mathcal{M}$. The argument proceeds in five parts: small-ball asymptotics on $\mathcal{M}$, a cross-mass identity, a sandwich for the k-NN radius, Bernstein concentration of the degrees, and a density-ratio bound through the KL gap.

We begin with small-ball asymptotics. Since $\mathcal{M}$ is compact, there exist constants $r_0, c_1, c_2 > 0$ (depending only on the Riemannian geometry of $\mathcal{M}$) such that for all $x \in \mathcal{M}$ and $0 < r \le r_0$:

$$c_1 r^d \;\le\; \mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(x,r)) \;\le\; c_2 r^d,$$

and moreover, uniformly in x:

$$\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(x,r)) = \omega_d r^d\big(1+O(r^2)\big).$$

Consequently, for the densities f and $\hat{f}$ with respect to $\mathrm{vol}_{\mathcal{M}}$:

$$p_1(B_{\mathcal{M}}(x,r)) = f(x)\,\omega_d r^d + O(r^{d+2}), \qquad \hat{p}_1(B_{\mathcal{M}}(x,r)) = \hat{f}(x)\,\omega_d r^d + O(r^{d+2}), \tag{14}$$

uniformly in $x \in \mathcal{M}$ for $r \le r_0$. In particular, $m\omega_d r^d/2 \le p_1(B_{\mathcal{M}}(x,r)) \le 2M\omega_d r^d$ for all x and $r \le r_0$.

We next establish a cross-mass identity. For $q > 0$ sufficiently small, define the quantile radius $r_q(x)$ by

$$p_1\big(B_{\mathcal{M}}(x, r_q(x))\big) = q.$$

By equation 14 and the density bounds $m \le f \le M$, for $q \le 2\gamma$ with γ small enough this is well-defined for every $x \in \mathcal{M}$, and

$$\left(\frac{q}{2M\omega_d}\right)^{1/d} \le r_q(x) \le \left(\frac{2q}{m\omega_d}\right)^{1/d}.$$

In particular, $r_q(x) \le Cq^{1/d}$.

Define the *cross-mass function*

$$h_q(y) \;:=\; \hat{p}_1\big(\{x \in \mathcal{M} : y \in B_{\mathcal{M}}(x,\, r_q(x))\}\big).$$

We claim that

$$\sup_{y\in\mathcal{M}}\left|h_q(y) - q\,\frac{\hat{f}(y)}{f(y)}\right| \;\le\; C\,q^{1+1/d}. \tag{15}$$

To prove equation 15, fix y and set $A_q(y) = \{x : y \in B_{\mathcal{M}}(x, r_q(x))\}$, so $h_q(y) = \hat{p}_1(A_q(y))$. If $x \in A_q(y)$, then $d_{\mathcal{M}}(x,y) \le r_q(x) \le Cq^{1/d}$. By Lipschitz continuity of f:

$$f(x) = f(y) + O\Big(q^{1/d}\Big).$$

Substituting into equation 14:

$$q = \omega_d f(x)\, r_q(x)^d + O\big(r_q(x)^{d+2}\big) = \omega_d f(y)\, r_q(x)^d\big(1+O(q^{1/d})\big),$$

so $r_q(x)^d = q/(\omega_d f(y)) \cdot (1 + O(q^{1/d}))$. Defining $a_q(y) = \big(q/(\omega_d f(y))\big)^{1/d}$, we obtain the geodesic ball sandwich

$$B_{\mathcal{M}}\Big(y,\ a_q(y)(1 - Cq^{1/d})\Big) \ \subset\ A_q(y) \ \subset\ B_{\mathcal{M}}\Big(y,\ a_q(y)(1 + Cq^{1/d})\Big).$$

Taking $\hat{p}_1$-measure of the outer set and using equation 14 for $\hat{p}_1$:

$$h_q(y) \ =\ \hat{f}(y)\,\omega_d\,a_q(y)^d + O\big(a_q(y)^{d+1}\big) \ =\ q\,\frac{\hat{f}(y)}{f(y)} + O\Big(q^{1+1/d}\Big),$$

uniformly in y. This proves equation 15.

We now derive a sandwich for the k-NN radius. Choose a constant $A > 0$ (to be fixed below) and define

$$q_\pm = \frac{k \pm A\sqrt{kL_N}}{N}.$$

If $k \geq C_0 L_N$ with $C_0 \geq 4A^2$, then $q_- > 0$. If $k \leq \gamma N$ for γ small enough, then q_+ is within the regime where the cross-mass identity equation 15 applies.

Let R_i denote the distance from $\hat{Y}_i$ to its k-th nearest target in $\{Y_1, \ldots, Y_N\}$. Conditionally on $\hat{Y}_i = x$, the count of targets in $B_{\mathcal{M}}(x, r_\pm(x))$ (where $r_\pm = r_{q_\pm}$) is $\mathrm{Binomial}(N, q_\pm)$, since the Y_j are i.i.d. from p_1 and $p_1(B_{\mathcal{M}}(x, r_{q_\pm}(x))) = q_\pm$. The Chernoff bound gives

$$\Pr\Big(R_i < r_-(\hat{Y}_i)\Big) \ \leq\ e^{-cA^2 L_N}, \qquad \text{and} \qquad \Pr\Big(R_i > r_+(\hat{Y}_i)\Big) \ \leq\ e^{-cA^2 L_N}.$$

Choosing A large enough so that $2Ne^{-cA^2 L_N} \leq \delta/3$, a union bound over $i = 1, \ldots, N$ gives an event E_1 with $\Pr(E_1) \geq 1 - \delta/3$ on which

$$r_-(\hat{Y}_i) \ \leq\ R_i \ \leq\ r_+(\hat{Y}_i), \qquad \forall\, i \in \{1, \ldots, N\}.$$

With this sandwich in hand, we turn to the degree concentration. On E_1, for every j: $Z_j^- \leq d_j(k) \leq Z_j^+$, where $Z_j^\pm = \sum_{i=1}^N \mathbf{1}\{Y_j \in B_{\mathcal{M}}(\hat{Y}_i, r_\pm(\hat{Y}_i))\}$. Conditionally on $Y_j = y$, the indicators are i.i.d. Bernoulli with success probability $h_{q_\pm}(y)$, so $Z_j^\pm \mid Y_j \sim \mathrm{Binomial}(N, h_{q_\pm}(Y_j))$.

Expectation. By equation 15 and $Nq_\pm = k \pm A\sqrt{kL_N}$:

$$\left|\mathbb{E}[Z_j^\pm \mid Y_j] - k\,\frac{\hat{f}(Y_j)}{f(Y_j)}\right| \ \leq\ C\sqrt{kL_N} + Ck\left(\frac{k}{N}\right)^{1/d}.$$

Variance. Since M/m bounds the density ratio, $h_{q_\pm}(y) \leq Ck/N$ for all y, so $\mathrm{Var}(Z_j^\pm \mid Y_j) \leq Nh_{q_\pm}(Y_j) \leq Ck$.

Bernstein. The Bernstein inequality gives, for $B > 0$:

$$\Pr\Big(|Z_j^\pm - \mathbb{E}[Z_j^\pm \mid Y_j]| > B\sqrt{kL_N}\Big) \ \leq\ 2e^{-cB^2 L_N}.$$

Choosing B so that $4Ne^{-cB^2 L_N} \leq 2\delta/3$ and union-bounding over $j = 1, \ldots, N$ and both signs $\pm$, we obtain an event E_2 with $\Pr(E_2) \geq 1 - 2\delta/3$ on which

$$\left|\frac{d_j(k)}{k} - \frac{\hat{f}(Y_j)}{f(Y_j)}\right| \ \leq\ C_1\sqrt{\frac{L_N}{k}} + C_2\left(\frac{k}{N}\right)^{1/d} \qquad \forall\, j \in \{1, \ldots, N\}.$$

It remains to control the density ratio $\hat{f}/f$ via the KL gap. By Pinsker's inequality: $\|p_1 - \hat{p}_1\|_{\mathrm{TV}} \leq \sqrt{\Delta_{\mathrm{KL}}/2}$, hence $\|f - \hat{f}\|_{L^1(\mathrm{vol}_{\mathcal{M}})} \leq \sqrt{2\Delta_{\mathrm{KL}}}$. Set $h = \hat{f} - f$ and $S = \|h\|_{L^\infty(\mathcal{M})}$. Choose $x_0 \in \mathcal{M}$ with $|h(x_0)| = S$. Since $|h|$ is Lipschitz with constant $2L$, for all x with $d_{\mathcal{M}}(x, x_0) \leq S/(4L)$ we have $|h(x)| \geq S/2$. Using $\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(x_0, r)) \geq c_1 r^d$:

$$\sqrt{2\Delta_{\mathrm{KL}}} \ \geq\ \|h\|_{L^1(\mathrm{vol}_{\mathcal{M}})} \ \geq\ \frac{S}{2} \cdot c_1\left(\frac{S}{4L}\right)^d \ =\ c\,S^{d+1}.$$

Therefore $S \le C\,\Delta_{\mathrm{KL}}^{1/(2(d+1))}$, and since $f \ge m$:

$$\sup_{y\in\mathcal{M}}\left|\frac{\hat f(y)}{f(y)}-1\right| = \sup_{y}\frac{|h(y)|}{f(y)} \le \frac{S}{m} \le C_3\,\Delta_{\mathrm{KL}}^{\frac{1}{2(d+1)}}.$$

Combining the two bounds yields the claim. On $E_1 \cap E_2$, which has probability $\ge 1-\delta$, the triangle inequality gives for every j:

$$\left|\frac{d_j(k)}{k}-1\right| \;\le\; \left|\frac{d_j(k)}{k}-\frac{\hat f(Y_j)}{f(Y_j)}\right| + \left|\frac{\hat f(Y_j)}{f(Y_j)}-1\right| \;\le\; C_1\sqrt{\frac{L_N}{k}} + C_2\left(\frac{k}{N}\right)^{1/d} + C_3\,\Delta_{\mathrm{KL}}^{\frac{1}{2(d+1)}}.$$

Taking the maximum over $j = 1,\dots,N$ yields equation 13. □

Corollary A.10. *Under Assumption A.1, for any $\delta \in (0,1)$ and any k such that $C_0\log(N/\delta) < k < \gamma N$, with probability at least $1-\delta$:*

$$\epsilon_k \;\le\; \Psi(k,\delta) \;:=\; \frac{a_1\,\log(N/\delta)}{k} + a_2\left(\frac{k}{N}\right)^{2/d} + a_3\,\Delta_{\mathrm{KL}}^{\frac{1}{d+1}}, \tag{16}$$

where $a_1 = 3C_1^2$, $a_2 = 3C_2^2$, and $a_3 = 3C_3^2$.

Proof. On the event of Theorem A.9, write $\alpha_j = |d_j(k)/k - 1|$ and let A, B, C denote the three terms on the right side of equation 13. Then $\alpha_j \le A+B+C$ for all j, and $(A+B+C)^2 \le 3(A^2+B^2+C^2)$. Hence

$$\frac{1}{N}\sum_j \alpha_j^2 \;\le\; (A+B+C)^2 \;\le\; a_1\,\frac{\log(N/\delta)}{k} + a_2\left(\frac{k}{N}\right)^{2/d} + a_3\,\Delta_{\mathrm{KL}}^{\frac{1}{d+1}}.$$

Lemma A.8 gives $\epsilon_k \le \frac{1}{N}\sum_j \alpha_j^2 \le \Psi(k,\delta)$. □

A.5 Main Threshold Certificate Theorem

Remark A.11. The computable bound

$$\Psi(k,\varepsilon) = \frac{a_1\log(N/\varepsilon)}{k} + a_2\left(\frac{k}{N}\right)^{2/d} + a_3\Delta_{\mathrm{KL}}^{1/(d+1)}$$

need not be monotone in k. It is used only to find a certified entry point. If the set

$$\mathcal{K}_{\mathrm{cert}}(\varepsilon) = \left\{k \in \{1,\dots,N\} : C_0\log(N/\varepsilon) < k < \gamma N,\ \Psi(k,\varepsilon) < \Delta_{\mathrm{KL}}\right\}$$

is non-empty, any element of it certifies the exact defect at that value of k, and Lemma A.2 then propagates the certificate to all larger graphs. Thus failure of the computable bound at a larger k reflects only looseness of Ψ, not failure of the exact defect.

Theorem A.12. *Let*

$$\Delta_\star := \Delta_{\mathrm{KL}} - a_3\,\Delta_{\mathrm{KL}}^{\frac{1}{d+1}}.$$

Assume $\Delta_\star > 0$. Define the candidate entry point

$$k_0 = \max\left\{\lfloor C_0\log(N/\varepsilon)\rfloor + 1,\ \left\lfloor\frac{2a_1\log(N/\varepsilon)}{\Delta_\star}\right\rfloor + 1\right\},$$

and

$$k_{\max} = \min\left\{\lceil\gamma N\rceil - 1,\ \left\lfloor\left(\frac{\Delta_\star}{2a_2}\right)^{d/2} N\right\rfloor\right\}.$$

If $k_0 \le k_{\max}$, then

$$\Psi(k_0,\varepsilon) < \Delta_{\mathrm{KL}}.$$

Consequently, with probability at least $1-\varepsilon$, for every integer $k \in \{k_0,\dots,N\}$ and every

$$\tau \;>\; \frac{\lambda_k + \eta\log(Nk)}{\Delta_{\mathrm{KL}} - \Psi(k_0,\varepsilon)},$$

the MASS plan satisfies

$$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) < \Delta_{\mathrm{KL}}$$

on the empirical target cloud.

Proof. By construction,

$$\Psi(k_0, \varepsilon) = \frac{a_1 \log(N/\varepsilon)}{k_0} + a_2 \left(\frac{k_0}{N}\right)^{2/d} + a_3 \, \Delta_{\mathrm{KL}}^{\frac{1}{d+1}}.$$

The lower bound in the definition of k_0 implies

$$\frac{a_1 \log(N/\varepsilon)}{k_0} < \frac{\Delta_\star}{2},$$

and the condition $k_0 \le k_{\max}$ implies

$$a_2 \left(\frac{k_0}{N}\right)^{2/d} \le \frac{\Delta_\star}{2}.$$

Hence

$$\Psi(k_0, \varepsilon) < \Delta_\star + a_3 \, \Delta_{\mathrm{KL}}^{\frac{1}{d+1}} = \Delta_{\mathrm{KL}}.$$

Also $C_0 \log(N/\varepsilon) < k_0 < \gamma N$, so Corollary A.10 applies at k_0. On its probability $1 - \varepsilon$ event,

$$\epsilon_{k_0} \le \Psi(k_0, \varepsilon) < \Delta_{\mathrm{KL}}.$$

For every $k \ge k_0$, Lemma A.2 gives $\epsilon_k \le \epsilon_{k_0} \le \Psi(k_0, \varepsilon)$. The stated condition on τ and Theorem A.5 therefore imply

$$D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b}) < \Delta_{\mathrm{KL}}.$$

□

A.6 A Setting Where the Certificate's Constants Are Explicit

The constants of Theorem A.12 are worst-case and not accessible in practice. In a setting simple enough to allow it they can nonetheless be pinned down completely, and that setting also lets us check the theorem's conclusion directly rather than through them.

Take p_1 uniform on the flat torus $\mathbb{T}^2 = [0,1)^2$, p_0 uniform, and as teacher the analytic map $(u, v) \mapsto (\phi(u), \phi(v))$ with $\phi(t) = t + \frac{\delta}{2\pi}\sin(2\pi t)$, so that Assumption A.1 holds exactly and teacher quality is controlled by the single parameter δ. Here $d = 2$ exactly; the curvature term in $\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(x, r)) = \omega_d r^d(1 + O(r^2))$ vanishes identically because the torus is flat; $\omega_d = c_1 = c_2 = \pi$; and the target-cloud diameter is $D = \sqrt{2}/2$. At $\delta = 0.5$ the density bounds are $m = 0.444$ and $M = 4.0$, the Lipschitz constant is $L = 24.8$, and the baseline quality levels are $\Delta_{\mathrm{KL}} = -2\log\big((1 + \sqrt{1 - \delta^2})/2\big) = 0.139$ and $\Delta_W = 0.080$. The three quantities that are inaccessible in general — intrinsic dimension, density, and curvature — are therefore all explicit in this setting.

The conclusion can be checked without them. With uniform marginals the k-NN graph carries a plan with exactly the right target marginal as soon as it admits a perfect matching, so the graph defect vanishes at that threshold. At $N = 2{,}000$ we measure the threshold at $k = 82$: beyond it $\epsilon_k = 0$, and the hypotheses of Theorems A.5 and A.7 hold exactly rather than sufficiently. This does not make the constants computable at image scale, but it shows that the mechanism the theorem describes — a defect that falls to zero once the graph is rich enough — is real and locatable when the geometry is known.

B Semi-Relaxed Robust Sinkhorn Details

Algorithm 3 reproduces the dense semi-relaxed Sinkhorn updates in the style of prior presentations of SR-Sinkhorn (Le et al., 2021; Fukunaga & Kasai, 2022). Algorithm 4 is the version aligned with our implementation: it operates in log scale, uses the sparse support induced by the kNN graph, zeros disconnected target columns, and row-normalizes the final plan before stochastic sampling.

Algorithm 3 Semi-Relaxed Robust Sinkhorn (Le et al., 2021; Fukunaga & Kasai, 2022)

Input: cost $\mathbf{C}$, marginals $\mathbf{a}, \mathbf{b}$, parameters η, τ, iterations n_{iter}
Initialize $\mathbf{u}^{(0)} \leftarrow \mathbf{0}$, $\mathbf{v}^{(0)} \leftarrow \mathbf{0}$
for $\ell = 0$ **to** $n_{\text{iter}} - 1$ **do**
  $\mathbf{T}^{(\ell)} \leftarrow \operatorname{diag}(e^{\mathbf{u}^{(\ell)}/\eta})\, e^{-\mathbf{C}/\eta} \operatorname{diag}(e^{\mathbf{v}^{(\ell)}/\eta})$
  if ℓ even **then**
    $\mathbf{u}^{(\ell+1)} \leftarrow \frac{\tau\eta}{\tau+\eta}\left(\frac{\mathbf{u}^{(\ell)}}{\eta} + \log \mathbf{a} - \log(\mathbf{T}^{(\ell)}\mathbf{1})\right)$
    $\mathbf{v}^{(\ell+1)} \leftarrow \mathbf{v}^{(\ell)}$
  else
    $\mathbf{u}^{(\ell+1)} \leftarrow \mathbf{u}^{(\ell)}$
    $\mathbf{v}^{(\ell+1)} \leftarrow \mathbf{v}^{(\ell)} + \eta\left(\log \mathbf{b} - \log((\mathbf{T}^{(\ell)})^\top \mathbf{1})\right)$
  end if
end for
$\mathbf{T} \leftarrow \operatorname{diag}(e^{\mathbf{u}^{(n_{\text{iter}})}/\eta})\, e^{-\mathbf{C}/\eta} \operatorname{diag}(e^{\mathbf{v}^{(n_{\text{iter}})}/\eta})$
Return: $\mathbf{T}$

Algorithm 4 Log-Scale Semi-Relaxed Robust Sinkhorn (MASS implementation)

Input: cost $\mathbf{C}$, $\mathbf{a}, \mathbf{b}$, η, τ, n_{iter}
$\mathbf{f} \leftarrow \mathbf{0}$, $\mathbf{g} \leftarrow \mathbf{0}$, $\mathbf{K} \leftarrow e^{-\mathbf{C}/\eta}$
for $\ell = 0$ **to** $n_{\text{iter}} - 1$ **do**
  if ℓ even **then**
    $\mathbf{f} \leftarrow \log \mathbf{a} - \log(\mathbf{K} e^{\mathbf{g}})$
  else
    $\mathbf{g} \leftarrow \frac{\tau}{\tau+\eta}\left(\log \mathbf{b} - \log(\mathbf{K}^\top e^{\mathbf{f}})\right)$
    Zero entries of $\mathbf{g}$ for disconnected target columns
  end if
end for
$\mathbf{T} \leftarrow e^{\mathbf{f}} \odot \mathbf{K} \odot e^{\mathbf{g}}$
Row-normalize each nonempty row of $\mathbf{T}$ before sampling
Return: $\mathbf{T}$

C Experimental Details and Additional Experiments

This appendix gives the setup for every experiment in the paper together with the additional measurements referenced from Section 4. Appendix C.1 describes the CIFAR-10 setting, and Appendices C.2–C.6 then follow the order of the main text: what each component of the construction contributes, the graph defect measured at image scale, the flow geometry and sampling budget of the retrained models, the matched comparison against minibatch OT, and one-factor sensitivity together with plan reuse. The two synthetic studies come last: the 2D toy experiments that support the intuition of Section 4 (Appendix C.7) and the dependence on ambient dimension and sample size (Appendix C.8).

C.1 CIFAR-10 Implementation Details

All CIFAR-10 experiments were run on a single NVIDIA A100 PCIe 40 GB GPU. All training settings are taken unchanged from the official Mean Flow PyTorch reference implementation (Geng et al., 2025a;b); across FM, iMF, and retraining runs we only swap the loss and the coupling. Strong-teacher comparisons reuse the original Rectified Flow (Liu et al., 2023) training configuration unchanged, again differing only in the coupling.

On CIFAR-10 we build the MASS plan with k=200, semi-relaxed Sinkhorn parameters (τ, η)=(10,000, 50), and warm-up endpoints generated with m=50 Euler NFE. Neighbor search is carried out in pixel space, on the same 3,072-dimensional vectors that define the transport cost. The kNN graph is built with FAISS IVF and the sparse Sinkhorn updates run for up to 10,000 iterations; the full construction takes under one minute on a single A100 and is amortized once per warm-start checkpoint.

FID is evaluated with torchmetrics' `FrechetInceptionDistance` on 50,000 samples drawn with an Euler ODE sampler (10 NFE for FM/ReFM, 1 NFE for iMF/MASS ReiMF). The intrinsic floor quoted for Table 1 is computed separately with pytorch-fid on the CIFAR-10 train set: uniform with-replacement sampling gives FID $= 0.528 \pm 0.003$ at N=50,000 (mean and std over 5 seeds), 1.06 at N=25,000, and 2.67 at N=10,000, consistent with the $\sim N^{-1}$ Gaussian-estimation rate. The script that reproduces it is in the supplementary code.

C.2 Component Controls

MASS combines three ingredients: the teacher proposes the candidate neighborhoods, those neighborhoods are anchored in real data, and Sinkhorn reweights them. This experiment asks which of the three the downstream gain actually comes from.

We therefore build four alternative plans, each removing or replacing one ingredient while everything else is held fixed — the same FM-100k teacher, the same k=200 neighborhoods, and the same 20k-step continuation. Uniform kNN draws a neighbor uniformly instead of solving for weights, and so drops the reweighting. Nearest-neighbor correction keeps a single candidate per source, and so drops the breadth as well. Shuffling the teacher endpoints across sources permutes the rows of the plan: the set of targets it produces is untouched, and only the correspondence between a source and its target is destroyed. Random support replaces the teacher's neighborhoods with random ones and runs the identical solver, keeping Sinkhorn while removing the model alignment. Each plan is then scored twice, once as a target set and once as a coupling (Table 3).

The two scores disagree, and that disagreement is the result. A construction can reproduce the target set of MASS almost exactly and still give up most of the retraining gain, and one of them scores better than MASS as a target set while retraining worse. A good target marginal is therefore not what makes the plan useful. Sinkhorn fixes the marginal, real-data anchoring keeps the endpoints on the data, and model alignment supplies the correspondence between source and target — which the target set cannot express and only a retrained model reveals.

Table 3: CIFAR-10 component controls. All plans use the FM-100k teacher and k=200 neighborhoods. Induced FID scores the target set the fixed plan produces; retrained FID scores generated samples after 20k further continuation steps, at 10 NFE. Lower is better in both columns.

Construction	Induced FID	Retrained FID
MASS	0.64	17.46
Uniform kNN	16.85	28.16
Nearest-neighbor correction	21.75	28.62
Shuffled teacher endpoints	0.63	20.30
Random support + Sinkhorn	0.40	18.14

C.3 Marginal Diagnostics on the CIFAR-10 Pixel Graph

Table 1 compares the plans by induced-target FID alone. For the same plans we also measured the coverage of the k-NN graph and the defect of the plan's target marginal (Table 4). Both improve steadily with k and track induced FID across the whole grid, but they do not track the retrained model: past k=200 both keep improving while retraining does not. These diagnostics are therefore useful for sizing the graph, and no substitute for a retraining comparison.

Table 4: Marginal diagnostics on the CIFAR-10 pixel-space plans: FM-100k endpoints, N=50,000, ambient dimension 3,072, η=50, τ=10^4. Coverage is the fraction of target points the k-NN graph reaches; $D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b})$ is the marginal defect the solved plan achieves; induced FID scores the target set it produces. Lower is better in the last two columns.

k	Coverage	$D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b})$	Induced FID
50	0.790	0.2980	1.94
100	0.866	0.1750	1.03
200	0.924	0.0915	0.62
500	0.970	0.0339	0.50

C.4 Flow Geometry and Sampling Budget

We re-evaluate the four 400k models of Figure 4(b) across the sampling budget, and add a measure of how straight each model's trajectories are: the per-dimension residual

$$S = \int_0^1 \mathbb{E}\left\| v_\theta(x_t, t) - (x_1 - x_0) \right\|^2 dt,$$

on 2,048 trajectories at 100-step Euler (Table 5).

Straightness alone does not explain the ordering: ReiMF is the second-straightest model and the worst generator, because regressing onto teacher-generated pseudo-targets straightens the paths while moving their endpoints off the data. MASS is straighter still and keeps its endpoints in the empirical sample, and that combination is what produces the one-step result. Its margin over Dec-iMF narrows from 3.01 FID at one step to 0.32 at eight without reversing, so the gain is a few-step gain rather than a uniformly better model.

Table 5: The 400k models of Figure 4(b), re-evaluated under one common sampler; the 1-NFE column therefore differs slightly from that figure. Lower is better throughout.

Model	Straightness	1 NFE	2 NFE	4 NFE	8 NFE
MASS ReiMF	**0.0394**	**12.60**	**9.35**	**7.83**	**7.74**
Dec-iMF	0.0476	15.61	10.94	8.18	8.06
ReiMF	0.0408	24.16	20.92	19.47	18.44
iMF	0.0477	21.41	15.26	13.17	12.88

C.5 Matched Comparison Against Minibatch OT

We continue from FM-100k changing only the coupling, with minibatch OT re-solving an exact assignment on every fresh draw of B images (Table 6). The budgets match: the assignment costs 0.4 ms at B=32 and 8.7 ms at B=256, and MASS solves nothing during training.

Neither coupling dominates, and the reason is structural: a within-batch permutation improves the pairing immediately but is redrawn every step, whereas a plan solved once against the whole dataset starts slower and goes further. What separates the two in practice is memory rather than time: minibatch OT allocates only a transient $B \times B$ cost matrix, 256 KB at B=256, whereas MASS keeps one 241 MB plan resident throughout training. Exact global OT we could not run at all, its dense cost matrix alone being 10 GB in fp32 at N=50,000.

Table 6: Matched CIFAR-10 continuation from FM-100k, 50,000 samples at 1 NFE; columns are total training steps. Lower is better.

Coupling	150k	200k
Independent	31.69	19.93
Minibatch OT, B=32	28.41	18.88
Minibatch OT, B=256	24.71	16.73
MASS, k=500	25.79	16.31

C.6 Endpoint Budget, Sensitivity, and Plan Reuse

Endpoint generation versus evaluation. Two budgets are easily confused: the endpoint budget m is spent once, to build the plan, and the sampling NFE is spent every time the retrained model generates. Table 7 varies the first and scores the plan alone; Table 8 varies the second, after 20k continuation steps from FM-100k.

Induced FID improves monotonically in both k and m, and the gain from a larger k shrinks as m drops: at a low endpoint budget the bottleneck moves from the reweighting to the endpoints themselves, which k cannot compensate for. The best k does not move with the sampling budget. The 1- and 2-NFE columns of that second table separate nothing, since a model trained on the FM objective is not a usable generator at one or two Euler steps.

Table 7: Induced-target FID as the endpoint budget and neighborhood size vary. These are properties of the fixed plan, not FIDs of a retrained model.

Endpoint NFE m	$k = 50$	$k = 100$	$k = 200$	$k = 500$
1	61.82	47.38	36.16	25.27
2	40.90	32.16	25.10	17.71
5	15.34	11.68	8.51	4.94
10	6.49	4.44	2.69	1.13
50	1.94	1.03	0.62	0.50

Table 8: Retrained FM FID after 20k continuation steps from FM-100k, with 50,000 Euler samples at each evaluation budget.

k	1 NFE	2 NFE	4 NFE	10 NFE
50	359.2	130.1	43.15	18.28
200	357.0	129.3	41.53	17.46
500	364.1	132.5	41.51	17.44

One-factor sensitivity. We change one parameter at a time from $(k, m, \eta, \tau) = (200, 50, 50, 10^4)$, each run continuing 20k steps from FM-100k and evaluated on 50,000 samples at 10 Euler NFE (Table 9).

Three of the four axes are flat and one is sharp, so the method does not live on a tuned point: k saturates by 200, a $20\times$ range of η moves FID by 0.26, and τ costs 1.4 FID only when loosened far enough to let the graph defect roughly double. The endpoint budget is the exception, and the one setting worth choosing carefully. These observations describe the tested checkpoint and budget, not a universal prescription.

Table 9: CIFAR-10 sensitivity around the reference setting. Unlisted parameters stay fixed.

Setting	Retrained FID
Reference	17.46
$k = 50$	18.28
$k = 500$	17.44
$m = 10$	18.69
$m = 2$	31.16
$\eta = 10$	17.43
$\eta = 200$	17.69
$\tau = 10^3$	18.82
$\tau = 10^5$	17.35

Rebuilding the fixed plan. At the 200k checkpoint — already 100k steps into retraining on a plan built once from FM-100k — we continue two ways: keeping the original plan, and rebuilding it from the current student with k and the Sinkhorn parameters unchanged (Table 10).

Rebuilding gives no consistent gain, so the plan is not what goes stale. The margin over independent coupling does narrow over the full horizon, from 3.62 FID at 200k to 2.94 at 400k, but rebuilding recovers none of it: what is outgrown is the coarse geometric prior itself rather than the age of the plan built from it. Cost is not the issue either — a refresh is one generation pass plus the $O(kN)$ construction, under a minute — so this supports reuse here without ruling out benefits from refresh on other schedules.

Table 10: FID under fixed-plan reuse and refresh from the same 200k student.

Total steps	Plan kept	Plan rebuilt at 200k
200k	16.31	16.31
210k	15.71	15.81
220k	15.47	15.45
230k	15.28	15.24
240k	14.91	15.09

C.7 Additional 2D Toy Experiments

This section collects the 2D toy experiments that support the intuition in the main text (Figure 5).

They also fix the 2D operating point $(k, \tau, \eta, m) = (200, 100, 0.05, 10)$, which is robust on three knobs. A value $\tau \geq 10$ activates the regime that Theorem 3.1 requires (Table 11). The entropic parameter η controls within-neighborhood sharpness while keeping $D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b}) \leq 2 \times 10^{-4}$ across three orders of magnitude (Table 12). And $m \geq 2$ endpoint NFE suffices, with no further gain past m=5 (Figure 7b), so the warm-up generator can be very cheap. This schedule does not transfer unchanged to CIFAR-10, where the endpoint budget is the sharp axis (Appendix C.6).

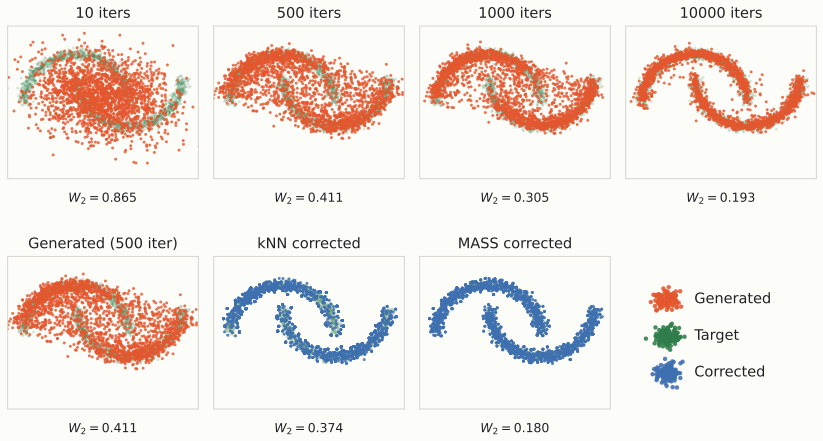

Figure 5: Top row: FM samples after $10, 500, 1000, 10000$ training iterations. Bottom row: 500-step samples with k=1 NN correction and MASS (k=200). Each panel reports W_2 to the true moons.

The source distribution p_0 is the standard 2D Gaussian and the target distribution p_1 is the standard two-moons distribution with $N = 2000$ samples. We train a small MLP velocity field with the standard Conditional Flow Matching objective equation 1 and independent coupling, taking checkpoints at 10, 500, 1000, and 10,000 iterations. For the correction comparison we start from the 500-iteration checkpoint, generate endpoints with a 10-step Euler integrator, and apply either k=1 nearest-neighbor replacement or MASS with $k = 200$ on the same kNN graph. We report W_2 between the resulting transported empirical target and the true moons cloud.

The setup mirrors two findings already seen on CIFAR-10. The early-stage model becomes geometrically informative well before it becomes a good generator; by 500 iterations the generated endpoints already concentrate near the two crescents, even though substantial fidelity gaps remain that only close at 1000 and 10,000 iterations. And k=1 nearest-neighbor replacement improves W_2 from 0.411 (raw generated samples) to 0.374, while MASS at $k = 200$ improves it further to 0.181, comparable to the much later 1000-iteration model. The same pattern holds as in Section 4.1, where nearest-neighbor snapping helps but distorts the marginal, and the MASS reweighting on a wider candidate set is what closes the gap.

To check that the qualitative story is robust, we re-ran the moons experiment in a separate self-contained implementation using a different random seed, moons noise level, and source-data normalization. Absolute W_2 values therefore differ from those above, but the ranking and trends are stable. With a deliberately weak FM-100 teacher (raw $W_2 = 0.324$ to a fresh moons sample, k=1 NN snap $W_2 = 0.263$), MASS at $\tau = 100$, $k = 200$ reaches $W_2 = 0.067 \pm 0.006$ averaged over 10 stochastic draws from the row-stochastic plan, which is below even the $W_2 = 0.176$ achieved by training the same model for 10,000 iterations. Table 11 reports the full τ-sweep at this weak teacher. The marginal-quality gain depends critically on τ being large enough; $\tau = 1$ gives only modest improvement, while $\tau \geq 10$ recovers the strong regime that the theorems require.

Figure 6(a) directly measures the quantity that Theorem A.4 bounds, namely $D_{\mathrm{KL}}(\mathbf{b}^k \| \mathbf{b})$ along the MASS plan, and overlays the Lemma A.8 degree-imbalance upper bound. The measured KL lies well below the bound, since Sinkhorn is strictly tighter than the uniform-graph plan, and both decrease in k as the theory predicts.

Table 11: 2D moons τ/k-sweep with FM-100 as weak teacher (lower is better). W_2 is against a fresh N=2,000 moons sample, averaged over 10 stochastic draws from the plan; $D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$ is the empirical defect that Theorem 3.3 controls.

		MASS at neighborhood size k					
τ	metric	k=25	k=50	k=100	k=200	k=500	(k=1 NN)
1	W_2	0.283	0.298	0.310	0.313	0.256	0.263
10	W_2	0.218	0.199	0.165	0.123	0.081	
100	W_2	0.191	0.155	0.109	0.067	0.058	
1	$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$	$1.2{\cdot}10^{-1}$	$9.8{\cdot}10^{-2}$	$9.4{\cdot}10^{-2}$	$8.8{\cdot}10^{-2}$	$4.9{\cdot}10^{-2}$	
10	$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$	$2.0{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	$5.9{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	
100	$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$	$7.6{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$	$2.1{\cdot}10^{-4}$	$8.6{\cdot}10^{-5}$	$2.0{\cdot}10^{-5}$	

To probe the high-dimensional caution of Remark A.11 on real data, we re-run the same defect computation on N=25,000 CIFAR-10 train Inception features (ambient dim 2048), using 25,000 independent same-distribution queries (clean) and queries perturbed by Gaussian noise of std 1.5 times the feature std (Figure 6(b)). Defect is again monotonically decreasing in k, and a finite-d power-law fit recovers an effective intrinsic dimension $d_{\mathrm{eff}} \approx 100$, much smaller than the ambient 2048. The residual defect at large k is dominated by hubness in high-dim NN graphs, an effect not captured by the $(k/N)^{2/d}$ term, and Sinkhorn's actual KL is much smaller than this loose upper bound. Noise-perturbed queries produce *lower* defect than clean queries, opposite the theory's $\Delta_{\mathrm{KL}}^{1/(d+1)}$ direction, because off-manifold Gaussian noise makes each query roughly equidistant from all features by the curse-of-dimensionality and the theory's manifold-support assumption is violated. Together the two panels support Remark A.11's reading of the bound as a structural guide rather than a tight image-scale prediction.

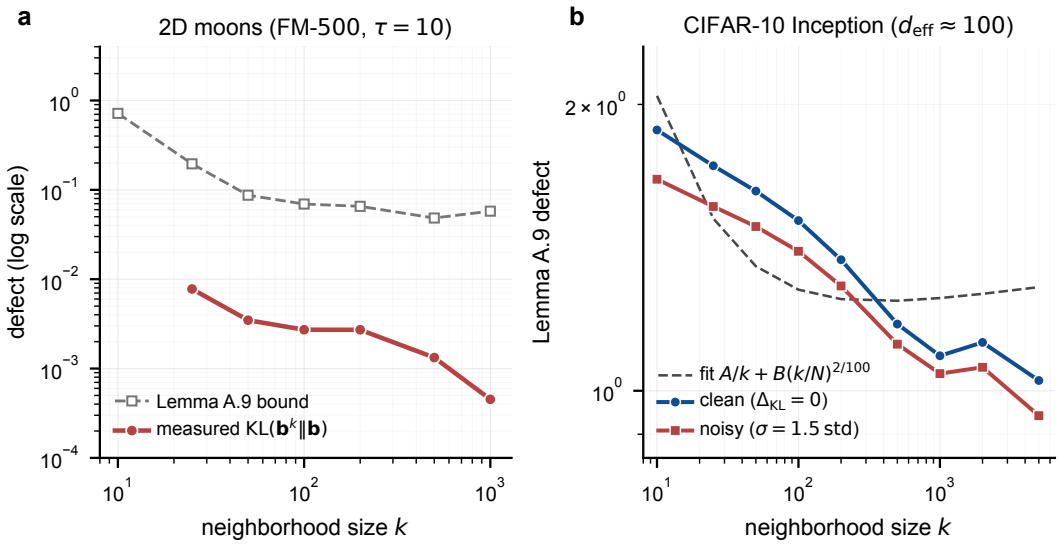

Figure 6: Theory and experiment bridge. (a) 2D moons (FM-500, τ=10): measured $D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$ along the MASS plan vs the Lemma A.8 bound. (b) CIFAR-10 Inception features (N=25,000, ambient dim 2048): defect for clean queries and heavy off-manifold noise, with finite-d fit yielding $d_{\mathrm{eff}} \approx 100$.

We further compare the MASS coupling against the natural OT baselines on the same FM-100 weak teacher, evaluated by the *transport cost* $\frac{1}{N}\sum_i \|x_0^i - x_1^{\pi(i)}\|^2$, the quantity that OT itself minimizes (lower is better, and W_2 on the marginal alone is uninformative for permutation-based methods, since it is constant for any permutation of the training set). Independent coupling has transport cost 3.95. With the FM-100 teacher, k=1 NN snap reaches 0.254, but its target marginal is severely

distorted ($D_{\mathrm{KL}}(\text{NN marg}\|\mathbf{b}) = 1.76$). Minibatch OT with batch sizes $B \in \{64, 256, 1024\}$ achieves transport cost $\{0.515, 0.375, 0.334\}$ respectively, and global OT on the full N=2000 dataset achieves 0.326. MASS at $k = 200$, $\tau = 100$, $m = 10$ achieves transport cost 0.336, within 3% of global OT. It improves on minibatch OT at $B = 64, 256$ and is close to $B = 1024$ (0.334), while keeping $D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b}) \approx 9 \times 10^{-5}$, four orders of magnitude below NN's marginal distortion. MASS therefore achieves global-OT-quality coupling and near-uniform target marginal at $O(kN)$ one-time cost. Figure 7(A) visualizes the full comparison.

The warm-up generator only provides a geometric hint, so a practical concern is whether MASS requires accurate model endpoints. We sweep the number of Euler steps $m \in \{1, 2, 5, 10, 20\}$ used to compute the endpoints $\hat{x}_{1,i}$ at fixed $k = 200, \tau = 100, \eta = 0.05$. Beyond a minimum threshold of $m = 2$ steps, transport cost is essentially flat $(0.342, 0.336, 0.336, 0.336)$ and the marginal defect changes by less than 30% ($D_{\mathrm{KL}} = 1.1{\times}10^{-4}, 9.1{\times}10^{-5}, 8.6{\times}10^{-5}, 8.4{\times}10^{-5}$). Only the degenerate $m = 1$ case fails, since a single Euler step from an untrained model produces endpoints close to $x_{0,i}$ itself, so the kNN graph queries the wrong region of the dataset and the plan collapses. MASS endpoint generation can therefore use a very cheap $m = 2$ to 5 NFE schedule with no measurable cost in coupling quality, in line with the design intuition that the warm-up model only needs to provide a coarse geometric signal. Figure 7(B) shows this.

To complete the three-knob ablation, we sweep the Sinkhorn entropic parameter η at fixed $k = 200, \tau = 100, m = 10$ on the same FM-100 teacher. Table 12 reports the resulting transport cost, $D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$, and a third diagnostic, the *effective row support* $\mathrm{eff}_i = \exp(H(T^k_{\mathrm{row}}[i, :]))$ averaged over i, which counts how many targets each source effectively distributes mass across (an entropic analogue of the cardinality of the row support). The sweep shows that η acts as a within-neighborhood sharpness knob. Small η collapses each row to its few best targets (effective support 2.4 at $\eta = 10^{-3}$, transport cost 0.197), while large η spreads mass across the full $k = 200$ candidates (effective support 172 at $\eta = 1.0$, transport cost 0.437). Marginal quality $D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$ is small across the entire range (at most $2{\times}10^{-4}$), so η trades off transport cost against support breadth without breaking the marginal. The moons default $\eta = 0.05$ sits in the middle of this range (effective support 41); CIFAR-10 uses $\eta = 50$. Aggressive choices like $\eta = 0.005$ yield even lower transport cost in exchange for a sharper plan that is closer to the deterministic kNN limit, and may have higher per-epoch sampling variance during retraining.

Table 12: Entropic regularization η ablation on 2D moons (FM-100 teacher, $k = 200$, $\tau = 100$, $m = 10$). η controls a within-neighborhood sharpness/breadth tradeoff while leaving the marginal defect small.

η	0.001	0.005	0.01	0.05 (default)	0.1	0.5	1.0
transport cost	0.197	0.243	0.290	0.336	0.350	0.410	0.437
$D_{\mathrm{KL}}(\mathbf{b}^k\|\mathbf{b})$	$3.2{\cdot}10^{-6}$	$3.0{\cdot}10^{-5}$	$5.9{\cdot}10^{-5}$	$8.6{\cdot}10^{-5}$	$8.8{\cdot}10^{-5}$	$1.2{\cdot}10^{-4}$	$2.0{\cdot}10^{-4}$
effective row support	2.4	6.0	10.4	40.9	70.7	154.1	172.1

C.8 Dimension and Sample-Size Dependence

The plan is built once on N source points, so the density of the neighborhood graph should matter. To isolate that, we hold the data manifold fixed and vary only the ambient dimension: a 2D spiral is embedded in $\mathbb{R}^D$ for $D = 2, 8, 16$ through a random column-orthogonal projection, so that D acts only through the teacher, and we vary N alongside it. Each model takes a 1k-step FM warm-up and then 5k further steps in which only the coupling changes (Table 13).

MASS improves over independent pairing at $D = 2$ and 8, but at $D = 16$ with $N = 2{,}000$ the ranking reverses, and raising N to 8,000 restores it. N therefore has to grow with the ambient dimension for a fixed plan to stay worth reusing, which is consistent with CIFAR-10 at $N = 50{,}000$. What this exposes is a dependence on the density of the neighborhood graph; it does not by itself establish an intrinsic-dimension scaling law.

Table 13: Spiral W_2 to a fresh target sample as the ambient dimension and sample count vary, averaged over three seeds. The evaluation columns distinguish 10 from 1 NFE; lower is better.

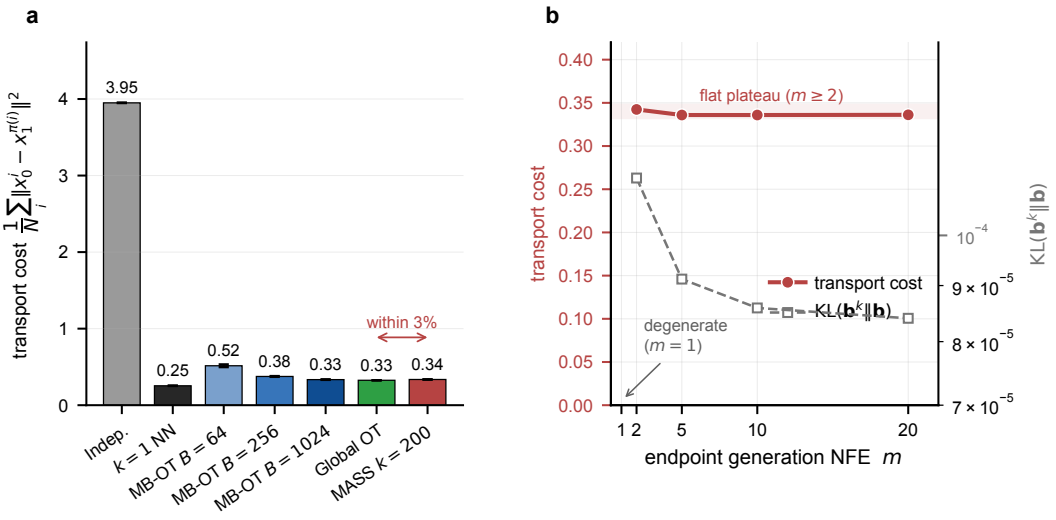

Figure 7: Coupling-construction comparison and NFE ablation on 2D moons (FM-100 weak teacher; lower is better). (A) Transport cost across coupling methods, with k=1 NN's marginal defect ($D_{\mathrm{KL}} = 1.76$) vs MASS's ($\sim 9{\times}10^{-5}$) noted in the body text. (B) MASS at k=200, τ=100 as m varies; the m=1 spike is a degenerate single-step regime.

			Independent		MASS	
D	N	Warm-up	10 NFE	1 NFE	10 NFE	1 NFE
2	2000	0.130	0.117	0.594	0.064	0.073
8	2000	0.153	0.123	0.581	0.101	0.263
16	2000	0.204	0.134	0.668	0.169	0.889
16	8000	0.206	0.141	0.673	0.127	0.579