COARSE-GRAINING OPTIMAL TRANSPORT FOR ROLLOUT-FREE POST-TRAINING OF FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching depends critically on the coupling used to pair source and target samples. Optimal transport (OT) couplings can reduce velocity conflict and straighten trajectories; but, in practice, challenges remain: the source comes from a continuous distribution, only a finite number of target samples data are accessible, and the neural network has finite capacity. For these reasons, making the coupling more OT-like may *not* consistently improve sampling quality, in particular in the low number-of-function-evaluation (NFE) regime. Motivated by this, we introduce ReCouple, a general framework that constructs coarse-grained OT couplings, to better understand rollout-free OT-based post-training. We also consider extensions of this basic framework: ReCouple+ extends this toward the continuous-source setting by increasing source resolution, converging to semi-discrete OT in the appropriate limit; and ReCouple++ further adds spatial regularization, motivated by the observation that straight trajectories need not correspond to spatially simple velocity fields. Empirical validation demonstrates the effectiveness of these methods, and it reveals an NFE-dependent interplay between coupling optimality, neural realizability, and generation quality.

1 INTRODUCTION

Flow matching (FM) (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023) provides an elegant generative modeling framework based on dynamical transport between probability distributions. Given a prescribed probability path from a simple source distribution to the target distribution, FM learns the associated velocity field with a neural network and generates samples by integrating the resulting ordinary differential equation (ODE). However, accurately resolving the learned flow often needs a large number of function evaluations (NFEs), requiring many neural network evaluations, particularly so when the flow trajectories are "curved."

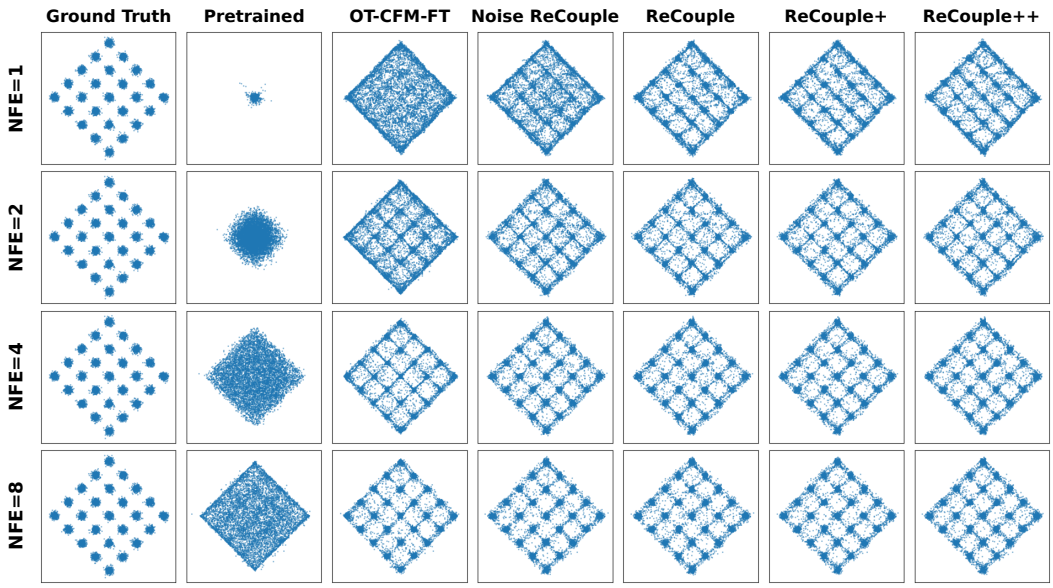
The "curvature" or "straightness" of a flow is determined by the coupling between the source and target, which we make precise here. For probability measures $\nu, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ (i.e., those with finite second moments), we let $\Pi(\nu, \mu)$ denote the set of couplings with marginals ν and μ. Given a coupling distribution $\gamma \in \Pi(\nu, \mu)$, consider the source-target pair $(Z, Y) \sim \gamma$, and the probability path $X_t = (1-t)Z + tY$, $t \in [0, 1]$. Conditioned on the pair (Z, Y), the path is a straight line, and $\dot{X}_t = Y - Z$. The regression target in FM is the velocity field $v_t^*(x; \gamma) = \mathbb{E}_\gamma[\dot{X}_t | X_t = x]$. Thus, the target depends not only on the probability path, but also on the choice of the coupling.

A simple choice is a random coupling that pairs Z and Y independently. Although each conditional path is straight, the marginal vector field averages velocities from different source-target pairs, often resulting in curved flows. To see this, the interpolation X_t induces a distribution of endpoint velocities $U = X_1 - X_0$ conditional on the intermediate state $X_t = x$, whose mean is the marginal velocity $v_t^*(x; \gamma)$. Averaging conflicting velocities locally in space produces a velocity field whose direction can vary along trajectories over time, resulting in curved paths and larger discretization error at low NFEs (Liu et al., 2022). When the conditional velocity distribution is concentrated near a single velocity, this local ambiguity is reduced and the induced flow is easier to integrate.

Optimal transport (OT) instead chooses the minimum-cost pairing, giving the OT plan $\gamma_{\mathrm{OT}} \in \arg\min_{\gamma \in \Pi(\nu,\mu)} \mathbb{E}_\gamma[\|Z - Y\|^2]$ for quadratic cost (also equivalent to maximizing endpoint alignment). Under quadratic OT, the induced displacement interpolation is a Wasserstein geodesic, so

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.

Figure 1: Generated samples on the 2D multimodal task across methods and NFEs. Post-training using the ReCouple family improves recovery of the target modes, particularly at low NFEs.



particles move with constant velocity along straight characteristics. Global OT over a large dataset is computationally costly, with standard solvers scaling cubically in sample size. In practice, batchwise OT solves OT only within each minibatch at each training iteration (Tong et al. (2023); Pooladian et al. (2023a)). However, this is suboptimal: each minibatch only covers a small region of the full distribution, so the batch-level couplings remain unstable across iterations and may result in curved flows. Due to these practical issues, OT-guided models are not widely used, and state-of-the-art models (e.g., Liu et al. (2022); Esser et al. (2024)) are typically trained on independent coupling pairs, often resulting in curved flows and requiring many NFEs to resolve the curvature.

OT can lead to straight flows realized by the Brenier map (Kornilov et al., 2024), but in practice we work with finite number of training/target samples whereas the source distribution is absolutely continuous (e.g., Gaussian), thus solving an OT problem which is semi-discrete in nature. However, both OT and semi-discrete OT (SDOT) are computationally costly (Peyré, 2025), while the Brenier map can be hard for a finite-capacity neural network to realize (Pierret et al., 2026). Moreover, even if we can approximate OT accurately, training naively on the OT pairs may not lead to a good few-step generator across all NFE regimes (Hui et al., 2025).

Despite these challenges, OT remains attractive for training and fine-tuning FM models. However, what are the caveats, and how far can this push a given pretrained FM model toward an improved few-step generator? This in turn exposes a basic question:

What makes a source-target coupling good for flow matching training in practice?

In this work, we introduce a controllable OT approximation framework to systematically understand the tradeoffs involved. To better expose the challenge and isolate approximation effects, we focus on post-training, and we default to a single post-training scheme (see Algorithm 2), avoiding adding various training "tricks" (Esser et al., 2024). These tricks can improve generative performance at the expense of higher computational cost, but they are often heuristic, task- and architecture-dependent (and thus less amenable to theory). Our scheme also does not require pretrained-model rollouts or teacher trajectories, unlike Reflow (Liu et al., 2022; Kim et al., 2025) and distillation-based schemes. It is also model-independent during coupling construction, unlike model-aligned approaches such as MAC (Lin et al., 2026) that use the current model to assess candidate pairs; see App. A for a discussion of related work. Throughout, we use explicit Euler as the default sampler at inference, so that NFE equals the number of sampling steps. Our focus is on improvements arising from the training procedure rather than from the choice of numerical solver.

Our main contributions are as follows.

- **ReCouple: an approximate OT scheme.** We introduce and study ReCouple, a divide-and-conquer approximation to OT based on structured blockwise assignment. This induces a natural *coarse-graining* of the transport problem: the block size B controls its resolution, ranging from projection-based rank matching at $B = 1$ to global empirical OT at $B = N$. We further provide mechanistic insights to understand ReCouple and its core components; see App. B.
- **ReCouple+: increasing the source resolution.** To increase source resolution, ReCouple+ replaces the one-to-one block assignments in ReCouple with pooled capacity-M assignments using MN source samples for N targets, separating geometric locality from continuous-source approximation. ReCouple+ thus serves as a controlled bridge between discrete OT and semi-discrete OT (SDOT), and we further show its convergence to SDOT in the appropriate limit (see App. D).
- **ReCouple++: introducing spatial regularization.** We observe empirically that approaching OT more closely needs not make the velocity field easier to learn. In particular, straighter transport trajectories can still correspond to spatially sharp Eulerian velocity fields with large Jacobians. Motivated by this distinction between trajectory straightness and spatial representational complexity, ReCouple++ augments ReCouple+ with spatial Jacobian regularization.
- **Empirical validations.** Empirical validations on multimodal 2D data and image generation tasks demonstrate the effectiveness of the ReCouple family compared to similar baselines. It also exposes the lesson that transport optimality and neural realizability need not improve sample quality together consistently across all NFE regimes. ReCouple++ turns the insights into an improved post-trained model by jointly controlling OT plan's approximation and spatial field complexity.

2 Flow Matching and Optimal Transport

2.1 Conditional Flow Matching (CFM)

Given a source-target pair $(Z, Y) \sim \gamma \in \Pi(\nu, \mu)$, define $X_t = (1-t)Z + tY$ and $U = Y - Z$.

CFM Training objective. The population CFM loss (Tong et al., 2023) is:

$$\mathcal{L}_{\mathrm{CFM}}(w;\gamma) = \mathbb{E}_{\substack{(Z,Y)\sim\gamma \\ t\sim\mathrm{Unif}[0,1]}} \left[\| v_t^w((1-t)Z + tY) - (Y - Z) \|^2 \right]. \tag{1}$$

We can decompose the loss into $\mathcal{L}_{\mathrm{CFM}}(w;\gamma) = \mathcal{E}(w;\gamma) + \mathcal{A}(\gamma)$, where $\mathcal{E}(w;\gamma) := \mathbb{E}_{(Z,Y)\sim\gamma,\, t\sim\mathrm{Unif}[0,1]} \| v_t^w(X_t) - v_t^*(X_t;\gamma) \|^2$ quantifies the field realization error and $\mathcal{A}(\gamma) := \mathbb{E}_{(Z,Y)\sim\gamma,\, t\sim\mathrm{Unif}[0,1]} [\mathrm{tr}\, \Sigma_t^\gamma(X_t)]$ quantifies the irreducible error. Here, $v_t^*(x;\gamma) := \mathbb{E}_\gamma[\dot{X}_t \mid X_t = x]$, the conditional covariance $\Sigma_t^\gamma(x) := \mathrm{Cov}_\gamma(\dot{X}_t \mid X_t = x)$ quantifies the disagreement among pathwise velocities passing through (t, x), with $\mathrm{tr}\, \Sigma_t^\gamma(x) = \mathbb{E}_\gamma[\| \dot{X}_t - v_t^*(x;\gamma) \|^2 \mid X_t = x]$. Thus, large Σ_t^γ corresponds to conflicting regression targets, and $\mathcal{A}(\gamma)$ reflects the ambiguity induced by the coupling. A good coupling should suppress $\mathcal{A}(\gamma)$, motivating schemes that reduce $\mathcal{A}(\gamma)$.

2.2 Semi-Discrete OT

Semi-discrete optimal transport (SDOT). Given an absolutely continuous (with respect to Lebesgue measure) source distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and a discrete empirical data distribution $\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$, where N is the sample size and $y_i \overset{\text{iid}}{\sim} \mu$ for some unknown target distribution, SDOT finds the optimal coupling (Peyré, 2025):

$$\gamma_N^\star \in \operatorname*{arg\,min}_{\gamma \in \Pi(\nu, \hat{\mu}_N)} \int \| z - y \|^2 \, \mathrm{d}\gamma(z, y). \tag{2}$$

Why SDOT is desirable. In practice, we are given access to only a finite number of training samples and we must learn the FM model using them. Thus, the CFM training operates in the data setting of SDOT, where the source distribution is continuous while the target distribution is discrete.

By Brenier's theorem (Villani, 2021), there exists a ν-a.e. unique OT map via a potential function ϕ:

$$T_N^\star = \nabla\phi, \qquad (T_N^\star)_\# \nu = \hat{\mu}_N, \qquad \gamma_N^\star = (\mathrm{Id}, T_N^\star)_\# \nu. \tag{3}$$

Algorithm 1 ReCouple

Require: target bank $y_{1:N}$; source bank $z_{1:N}$; block size $B \le N$ (B divides N); proposal count R

Add the identity pairing $\pi_0(i) = i$ to the candidate set.
for $r = 1, \dots, R$ **do**
  Draw $\theta_r \sim \mathrm{Unif}(\mathbb{S}^{d-1})$.
  Let $\sigma_r, \tau_r \in S_N$ sort $\{\langle \theta_r, y_i \rangle\}_{i=1}^N$ and $\{\langle \theta_r, z_j \rangle\}_{j=1}^N$, respectively
  Partition the ordered positions into $I_b = \{(b-1)B + 1, \dots, bB\}$, $b = 1, \dots, N/B$.
  for $b = 1, \dots, N/B$ **do**
    Compute

$$\rho_{r,b} \in \operatorname*{arg\,min}_{\rho \in \mathfrak{S}(I_b)} \sum_{k \in I_b} \left\| y_{\sigma_r(k)} - z_{\tau_r(\rho(k))} \right\|^2 .$$

  end for
  Define $\pi_r \in S_N$ by $\pi_r(\sigma_r(k)) = \tau_r(\rho_{r,b}(k))$ for $k \in I_b$.
  Evaluate $C_N(\pi_r) = \frac{1}{N} \sum_{i=1}^N \|y_i - z_{\pi_r(i)}\|^2$.
end for
Select $r^\star \in \arg\min_{r=0,\dots,R} C_N(\pi_r)$ and set $\widehat{\pi} = \pi_{r^\star}$.
return $\widehat{\gamma} = \frac{1}{N} \sum_{i=1}^N \delta_{(z_{\widehat{\pi}(i)}, y_i)}$.

Since $\widehat{\mu}_N$ is discrete, $T_N^\star$ is piecewise constant. In particular, there exist weights $\psi = (\psi_1, \dots, \psi_N)$ such that the Laguerre cells (see Section 6.3 in (Peyré, 2025))

$$L_i(\psi) = \left\{ z : \frac{1}{2}\|z - y_i\|^2 - \psi_i \le \frac{1}{2}\|z - y_j\|^2 - \psi_j, \ \forall j \right\} \tag{4}$$

satisfy $\nu(L_i) = 1/N$, and $T_N^\star(z) = y_i$ for ν-a.e. $z \in L_i$. The map T_N^* is discontinuous at the boundaries of each cell.

The following proposition highlights two favorable properties of the SDOT flow. Its deterministic coupling eliminates conditional velocity conflict, while its material acceleration, $D_t v_t := \partial_t v_t + (v_t \cdot \nabla) v_t$, vanishes along the displacement interpolation. Consequently, the velocity is constant along each characteristic, and a single Euler step from $t = 0$ to $t = 1$ recovers the SDOT map exactly.

Proposition 1 (Straightness of the SDOT flow)**.** *Let $T = T_N^\star$ be the Brenier map transporting ν to $\widehat{\mu}_N$, and define the affine interpolation $X_t = (1-t)Z + tT(Z)$, $\dot{X}_t = T(Z) - Z$, where $Z \sim \nu$. Let $v_t^*(x) := \mathbb{E}[\dot{X}_t \mid X_t = x]$ denote the corresponding marginal velocity field. Then, for every $0 \le t < 1$, $\mathrm{Cov}(\dot{X}_t \mid X_t) = 0$, and hence $v_t^*(X_t) = \dot{X}_t = T(Z) - Z$ a.s.. In particular, the velocity is constant along each characteristic,*

$$\frac{d}{dt} v_t^*(X_t) = \partial_t v_t^* + (v_t^* \cdot \nabla_x) v_t^* = 0, \tag{5}$$

wherever the derivatives exist.

Hence, the SDOT map provides a principled target for few-step generation in FM. Degradation at low NFEs thus arises from approximating the OT velocity field over $t \in [0, 1]$, rather than from intrinsic velocity conflict or curvature of the underlying flow trajectories. Approximate OT constructions (such as ReCouple and ReCouple+; see later subsections) can therefore be viewed as controlled approximations to this target map, allowing us to study how much of the OT geometry is needed at different NFE budgets. On the other hand, learning the discontinuous SDOT map with a spatially smooth velocity field can be difficult, thus at odds with the goal of exact one-step integration (see ReCouple++ in a later subsection).

3 RECOUPLE

3.1 RECOUPLE

How much OT is enough? We introduce a controllable framework to better understand this. The key to this framework is ReCouple (see Algo. 1 and Algo. 3 for details), an approximation

OT scheme that at a high level works as follows. We repeatedly draw a random direction, sort both source and target banks by projection onto the random direction, split the ordered lists into corresponding blocks of size B, solve the linear sum assignment (LSA) problem inside each block, and add the concatenated permutation into the candidate set. The final permutation is chosen as the one that realizes the minimal cost among the candidates. Each block-local LSA can be solved via the Hungarian or the modified Jonker–Volgenant algorithm (Crouse, 2016). Importantly:

> B controls the "coarse-graining scale" of the ordering induced by a 1D projection: after sorting source and target samples by their projected coordinates, consecutive groups of B samples define corresponding blocks within which the full OT is solved.

See App. B for additional insights. Throughout, S_N denotes the set of permutations of $\{1, \dots, N\}$.

ReCouple vs. SWGG and global empirical OT. ReCouple's plan can be viewed as a blockwise relaxation of the plan computed by minimizing SWGG (Mahey et al., 2023), with $B = 1$ recovers the SWGG's plan associated with the selected direction (see App. B for details), while $B = N$ recovers the global empirical OT coupling. By solving only local assignment problems induced by 1D random projections, ReCouple interpolates between the computational efficiency of SWGG-based plan computation and the geometric fidelity of global OT, reducing the per-proposal assignment cost from $O(N^3)$ to $O(NB^2)$. For nested blockings obtained by merging aligned finer blocks, increasing B enlarges the admissible set and cannot increase the blockwise OT cost.

3.2 ReCouple+

ReCouple+ is an enhanced version of ReCouple that increases the source resolution. In particular, the source bank is enlarged by a factor of M, in which case the LSA assignment problem extends to a capacity-M assignment problem (see Algo. 4 for details). Intuitively, increasing M reduces the finite-source approximation error.

This approximation scheme is not designed to be the most efficient solver for large-scale OT (Genevay et al., 2016; 2018). Rather, we exploit its explicit resolution parameters: B controls the scope of the admissible transport, while M controls the resolution of the continuous source approximation. Indeed, as B and M are increased toward the SDOT limit, the computational advantage of the blockwise approximation diminishes. Its purpose here is instead to provide an explicit and controllable family of couplings through which these different transport resolutions can be studied.

When does ReCouple+ approach SDOT? The parameters B and M control two distinct approximations to the SDOT problem. The block size B controls the *geometric resolution* of the coupling by restricting transport to corresponding projected-rank blocks, while the pooling factor M controls the *source resolution* by increasing the number of samples used to approximate the continuous source distribution. This distinction becomes clear at the two limits. When $B = N$, the block restriction disappears, and ReCouple+ solves the discrete OT problem between the pooled empirical source

$$\widehat{\nu}_{MN} = \frac{1}{MN} \sum_{j=1}^{MN} \delta_{Z_j}$$

and the empirical target $\widehat{\mu}_N$. As $M \to \infty$, $\widehat{\nu}_{MN}$ converges to ν in W_2 almost surely, and the ReCouple+ cost and optimal coupling converge to those of the SDOT problem between ν and $\widehat{\mu}_N$. In contrast, if $B < N$ is fixed and $M \to \infty$, the limit remains a block-restricted SDOT problem: increasing M resolves the continuous source more accurately but does not remove the geometric restriction imposed by B. At the limiting corner $B = N$ and $M \to \infty$, both approximations are removed and SDOT is recovered. See the theorem below, and App. D for more results and details.

Theorem 2 (Informal: Convergence of ReCouple+ to SDOT)**.** *Let $\widehat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{y_i}$ be given and suppose that $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous. Let $Z_i \overset{\text{iid}}{\sim} \nu$, and, for each $M \geq 1$, define $\widehat{\nu}_{MN} = \frac{1}{MN} \sum_{j=1}^{MN} \delta_{Z_j}$. Consider ReCouple+ with one global block, $B = N$, and pooling factor M. Let $\widehat{\gamma}^{+}_{N,M}$ denote the resulting empirical coupling. Then, $W_2(\widehat{\gamma}^{+}_{N,M}, \gamma_N^{\text{SDOT}}) \to 0$ almost surely, where γ_N^{SDOT} is the unique optimal coupling between ν and $\widehat{\mu}_N$.*

3.3 RECOUPLE++

Why ReCouple or ReCouple+ alone may not be optimal. While training on the ReCouple pairs can make the learned flow straighter across $t \in [0,1]$, it does not explicitly address the spatial variation of the velocity field (either in data space or latent space) during training. In particular, it does not account for nearby trajectories with substantially different velocities, which may be difficult for a finite-capacity model to distinguish. To account for this finite spatial resolution, we introduce a scale $\sigma > 0$ and define $W_t := X_t + \sigma\xi$, $\xi \sim \mathcal{N}(0, I)$, where ξ is independent of (Z, Y). For fixed t, define the time-local irreducibility error $\mathcal{A}_t(\gamma) := \mathbb{E}_{(Z,Y)\sim\gamma}\left[\operatorname{tr}\Sigma_t^{\gamma}(X_t)\right]$, and the Bayes risk:

$$\mathcal{D}_{\gamma,\sigma}(t) := \inf_f \mathbb{E}\left[\left\|\dot{X}_t - f(W_t)\right\|^2\right] = \mathbb{E}\left[\operatorname{tr}\operatorname{Cov}(\dot{X}_t \mid W_t)\right]. \tag{6}$$

Unlike $\mathcal{A}_t(\gamma)$, which only measures disagreement among velocities that meet at the same state, $\mathcal{D}_{\gamma,\sigma}(t)$ also regards sufficiently nearby trajectories as locally indistinguishable.

Let $J_\gamma(t)^2 := \mathbb{E}\left[\|\nabla_x v_t^*(X_t;\gamma)\|_F^2\right]$ for $t \in (0,1)$. Assuming that, for fixed t, X_t has a smooth positive density and $v_t^*(\cdot;\gamma)$ is sufficiently smooth with suitable integrability conditions, we have, as $\sigma \to 0$,

$$\mathcal{D}_{\gamma,\sigma}(t) = \mathcal{A}_t(\gamma) + \sigma^2 J_\gamma(t)^2 + \mathcal{O}(\sigma^4). \tag{7}$$

This suggests that suppressing the coupling ambiguity alone may not be enough, and adding spatial Jacobian regularization can be beneficial to improve ReCouple-based post-training. Therefore, we consider an enhanced version of ReCouple+, which we refer to as ReCouple++ (we refer to a similarly enhanced version of ReCouple as ReCouple-AS), whose training is done by minimizing

$$\mathcal{L}(w;\gamma) = \mathcal{L}_{CFM}(w;\gamma) + \lambda\, \mathbb{E}_{t\in\mathrm{Unif}[0,1]}\mathcal{R}_\sigma(w;t), \tag{8}$$

where

$$\mathcal{R}_\sigma(w;t) := \frac{1}{2\sigma^2}\mathbb{E}\left[\|v_t^w(X_t+\sigma\xi) - v_t^w(X_t)\|^2\right]. \tag{9}$$

Assuming that v_t^w is sufficiently smooth in x with suitable integrability conditions, a Taylor expansion yields, as $\sigma \to 0$,

$$\mathcal{R}_\sigma(w;t) = \frac{1}{2}\mathbb{E}\left[\|\nabla_x v_t^w(X_t)\|_F^2\right] + \mathcal{O}(\sigma^2). \tag{10}$$

Consequently, when v_t^w is close to the population field $v_t^*(\cdot;\gamma)$, $\mathcal{R}_\sigma(w;t) \approx \frac{1}{2}J_\gamma(t)^2$. We further anneal the penalty to zero so that the final training phase optimizes the unregularized CFM objective (see Alg. 5 in App. C for full details).

This issue becomes particularly relevant in the SDOT limit approached by ReCouple+ when $B = N$ and $M \to \infty$. Although SDOT displacement trajectories are exactly straight, the associated Eulerian velocity field is spatially sharp: within each transported Laguerre cell,

$$\nabla_x v_t^\star = -\frac{1}{1-t} I_d, \tag{11}$$

while globally stitching distinct cellwise branches requires Lipschitz scale at least $1/t$. Thus the SDOT velocity cannot remain uniformly spatially regular over time, motivating ReCouple++ to control the spatial variation of the learned field. See Theorem 9 in App. D.

Meanwhile, by Proposition 1, $\frac{d}{dt}v_t^*(X_t) = 0$. In other words, the ideal SDOT has perfectly straight characteristics, but the vector field is spatially sharp: it must distinguish many cells and needs to resolve increasingly steep transitions near the endpoints. By penalizing spatially high-variation modes, the annealed regularizer biases the learned velocity field toward smoother realizations during the initial phase of post-training. As the penalty is annealed to zero, the final phase optimizes the original CFM objective without an explicit spatial penalty.

Notably, (Hui et al., 2025) studied this phenomenon in 3D point cloud generation and mitigate it by injecting noise into the source samples. We instead develop an explicit block/pooling hierarchy and regularize the spatial variation of the neural field without modifying the underlying pair bank. We therefore also compare against an analogous noise-injection variant (Noise-ReCouple) which perturbs the ReCouple source endpoint as $z_\beta = \sqrt{1-\beta}\, z + \sqrt{\beta}\, \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, while keeping the target fixed.

4 EMPIRICAL VALIDATION

In this section, we evaluate post-training of FM models on the ReCouple family and related baselines on 2D multimodal generation and image generation tasks (CIFAR-10 and ImageNet-256). Details, ablation studies, and additional results are provided in App. E.

4.1 2D MULTIMODAL GENERATION

We consider unconditional generation for a 2D multimodal dataset (see App. E.1 for details). The source samples are from the standard Gaussian, while the target samples are from a 25-mode Gaussian mixture. Fig. 1 shows the samples generated by various baselines and the ReCouple schemes.

Table 1: Evaluation on the 2D multimodal task. We report the 2-Wasserstein distance (W_2, lower is better) at different NFEs, together with the neural acceleration energy $\mathcal{E}_{\mathrm{acc}}$ (lower is better).

Method	$W_2 \downarrow$				$\mathcal{E}_{\mathrm{acc}} \downarrow$
	NFE 1	NFE 2	NFE 4	NFE 8	
Pretrained	0.6103	0.4092	0.2119	0.1362	57.6858
PersistentIndependent-FT	0.5924	0.4145	0.2164	0.1378	65.4474
AlignFlow (Kong et al., 2026)	0.5234	0.3537	0.1880	0.1211	54.6471
OT-CFM-FT (Tong et al., 2023)	0.0996	0.0875	0.0848	0.0859	1.3101
SDOT (Mousavi-Hosseini et al., 2026)	0.1082	0.1092	0.1087	0.1087	**0.2628**
Noise-ReCouple (c.f. (Hui et al., 2025))	0.0895	0.0843	0.0849	0.0860	0.5316
ReCouple	0.0866	0.0853	0.0857	0.0862	0.4461
ReCouple-AS	0.0870	0.0848	0.0852	0.0858	0.3542
ReCouple+	**0.0859**	0.0845	0.0848	0.0853	0.4401
ReCouple++	0.0863	**0.0840**	**0.0843**	**0.0849**	0.3642

From Table 1, we see that ReCouple substantially improves over the pretrained and independent-FT baselines, while ReCouple+ achieves the best W_2 at NFE 1 and ReCouple++ performs best at NFE 2-8. The improvements are also reflected in the other distributional metrics: the ReCouple variants achieve substantially lower sW2 than OT-CFM and SDOT, while maintaining high recall and coverage (see Table 7-8 in App. E.1).

Spatial regularization provides a complementary control on realizability. It leads to consistently lower acceleration energy, while giving modest improvements at NFE 2-8 and little or no gain at NFE 1. Combining increased source resolution and spatial regularization in ReCouple++ gives the strongest overall W_2 performance beyond one step. Together, these results suggest that low-NFE quality depends on the interaction between the coupling, neural realizability of the induced velocity field, and numerical discretization, rather than on OT optimality alone.

4.2 CIFAR-10 IMAGE GENERATION

We evaluate the ReCouple family on the CIFAR-10 generation task in detail. We compare to the following baselines: (i) Fresh-independent FT: training on independent coupling, (ii) OT-CFM-FT: solving a minibatch-OT during each training step, (iii) SDOT baselines (Kong et al., 2026) and (Mousavi-Hosseini et al., 2026). We post-train ReCouple and the other similar baselines for 50k iterations using the same optimization setup during training.

Overall, ReCouple++ performs competitively at NFE-50 relative to similar baselines while providing a favorable balance across the various NFEs. ReCouple itself outperforms the noise injection strategy recently introduced by (Hui et al., 2025) (Noise-ReCouple) at NFE 1-2. Within the ReCouple family itself, ReCouple ($B = 2048$) tracks and shows similar trend in FID performance across the NFEs with ReCouple ($B = N$). Adding spatial regularization helps to boost the performance of both ReCouple and ReCouple+ across the NFEs. App. E provides ablation study on B, R and M, as well as other details.

Notably, ReCouple+ and ReCouple++ lead to higher FID at the ultra-low NFE regime compared to ReCouple and ReCouple-AS, in agreement with the discussion in subsection 3.3, but they shine in

Table 2: CIFAR-10 FID results with the Rectified Flow pretrained backbone. Lower is better. We fix $R = 32$ and $B = 2048$ for ReCouple, ReCouple-AS, ReCouple+, and ReCouple++. We use $M = 10$ for ReCouple+ and ReCouple++. Results are averaged over three seeds.

Method	NFE 1	NFE 2	NFE 4	NFE 8	NFE 10	NFE 50
RF (backbone)	357.341	165.085	50.014	16.821	13.123	5.180
Teacher Reflow (MSE)	11.417	**7.008**	**5.915**	**5.539**	**5.502**	**5.511**
One-step distillation (MSE)	**9.436**	128.435	149.356	163.575	167.244	182.325
Fresh-independent FT	356.823	165.896	50.278	16.853	13.110	5.033
OT-CFM-FT (Tong et al. (2023))	232.657	90.181	28.572	13.338	11.269	5.137
AlignFlow (Kong et al. (2026))	164.919	58.269	23.050	12.861	11.142	5.175
SD-FM (Mousavi-Hosseini et al. (2026))	159.871	58.982	23.201	12.727	11.003	5.087
Noise-ReCouple (cf. Hui et al. (2025))	132.991	47.773	22.852	16.109	14.956	10.313
ReCouple ($B = N$)	**102.110**	**41.520**	23.001	16.817	15.658	10.429
ReCouple	123.159	45.722	23.007	16.619	15.535	10.946
ReCouple-AS	122.554	45.333	**22.389**	16.141	15.047	10.518
ReCouple+	187.926	68.357	24.130	12.656	10.936	5.216
ReCouple++	187.080	67.905	23.797	**12.292**	**10.527**	**4.781**

Table 3: CIFAR-10 FID results with the OT-CFM pretrained backbone. Same setting as Table 2.

Method	NFE 1	NFE 2	NFE 4	NFE 8	NFE 10	NFE 50
OT-CFM (backbone)	230.321	86.640	27.312	12.709	10.723	4.902
Teacher Reflow (MSE)	11.196	**6.848**	**5.705**	**5.291**	**5.234**	**5.208**
One-step distillation (MSE)	**9.087**	137.009	159.789	177.066	181.694	201.234
Fresh-independent FT	351.241	155.466	47.210	15.845	12.286	4.634
OT-CFM-FT (Tong et al. (2023))	228.462	87.273	27.764	12.771	10.717	4.745
AlignFlow (Kong et al. (2026))	163.365	57.590	22.516	12.260	10.568	4.788
SD-FM (Mousavi-Hosseini et al. (2026))	157.251	57.958	22.743	12.243	10.546	4.779
Noise-ReCouple (cf. Hui et al. (2025))	132.299	47.230	22.280	15.596	14.496	10.001
ReCouple ($B = N$)	**101.005**	**40.877**	22.354	16.265	15.112	10.082
ReCouple	121.547	45.018	22.208	15.957	14.895	10.451
ReCouple-AS	120.577	44.297	**21.750**	15.480	14.431	10.085
ReCouple+	185.146	66.818	23.335	11.967	10.266	4.785
ReCouple++	184.021	65.635	22.878	**11.535**	**9.826**	**4.342**

the higher NFE regime. For a broader evaluation, we also include rollout-based training methods, which involve rolling out the teacher network using adaptive solver (using ≥ 100 NFEs), and are therefore do not compare fairly with the other baselines. The one-step distillation method specializes on one-step generation and performs poorly in the higher NFE regime.

4.3 ImageNet-256 Image Generation

We further evaluate the ReCouple family at scale by post-training a pretrained SiT-XL/2 backbone on ImageNet-256. We compare against the original backbone and minibatch OT under a matched sampling protocol. All results are computed from 50K generated samples using Euler integration and classifier-free guidance with scale 4.0. Additional results are provided in App. F.

Table 4 shows a clear NFE-dependent crossover. SD-FM obtains the lowest FID at NFE 1–2, whereas ReCouple++ obtains the lowest FID from NFE 5 onward. Within the ReCouple family, plain ReCouple performs best at NFE 1–2, while ReCouple++ performs best from NFE 5 onward. The results support our central observation: coupling choices that are strongest at 1-2 steps need not be strongest once the sampler has sufficient numerical resolution, while spatial regularization is particularly beneficial from NFE 5 onward.

Table 4: ImageNet-256 FID results using the pretrained SiT-XL/2 backbone. All metrics are computed using 50K generated samples with Euler integration. Lower is better; the best result at each NFE is shown in bold. ReCouple++ uses pooling factor $M = 2$.

Method	NFE 1	NFE 2	NFE 5	NFE 10	NFE 25	NFE 50
SiT-XL/2 FM	275.340	99.078	10.172	12.175	12.724	12.645
Minibatch OT (Tong et al. (2023))	253.866	90.371	7.845	9.036	9.437	9.360
SD-FM (Mousavi-Hosseini et al., 2026)	**239.573**	**80.828**	8.801	10.329	10.828	10.855
ReCouple	245.571	82.684	8.989	10.637	11.150	11.164
ReCouple-AS	246.547	88.086	7.275	7.051	7.432	7.539
ReCouple++ ($M = 2$)	246.249	87.645	**7.226**	**6.960**	**7.358**	**7.302**

Figure 2: Qualitative comparison on ImageNet-256 using pretrained SiT-XL/2 at NFE 25. Each row shows samples generated from the same initial noise and class condition using the SiT baseline, Minibatch OT, ReCouple, ReCouple-AS, and ReCouple++. This matched comparison highlights the differences in visual fidelity and semantic consistency induced by each sampling strategy.



5 Conclusion

We introduced the ReCouple family as a controllable framework for studying and designing endpoint couplings for post-training flow matching models, ranging from coarse blockwise transport to increasingly population-resolved and spatially regularized variants. Our findings suggest putting less emphasis on approximating OT as closely as possible. Instead, coupling design is better understood as a resolution matching problem. ReCouple controls the coarse-graining of the transport geometry, the pooling factor controls how densely that geometry is resolved from the continuous source, and Jacobian regularization controls the complexity of neural realization.

Our empirical results are consistent with this picture. ReCouple-based training gives competitive generative performance compared to similar baselines across low-NFE regime. Moreover, the ReCouple family gives an effective yet controllable approach to understand generative improvement across NFEs. The more interesting findings lie within the ReCouple family itself: refining the coupling improves transport fidelity, but may not lead to more substantial gains at the higher NFEs. Increasing source resolution favors higher NFEs at the expense of one-step performance, while spatial regularization mainly improves moderate-to-high NFEs. Taken together, these results suggest that the most effective coupling need not be the most OT-like, but rather one whose transport resolution is compatible with the data support, the neural realization, and the numerical budget.

AI USE STATEMENT

In this work, we conceived and carried out the original research questions, core ideas, proposed methodology, theoretical contributions, experimental design, and overall execution; generative AI tools were used only in an assistive capacity.

We used generative AI tools to assist with editing and polishing the text, checking mathematical arguments and derivations, and assisting with the implementation and verification of computational methods. We additionally used generative AI tools for improving clarity and presentation, brainstorming and refining research ideas, searching for and identifying potentially relevant literature, assisting with coding and research workflow, and reviewing the manuscript for consistency, clarity, and completeness.

We have reviewed all AI-assisted work. We have independently checked all mathematical arguments, derivations, and claims. We have also verified suggested references against the original sources. The code was inspected and tested, and all reported experimental results were verified. All substantive research decisions, including the formulation of the method, theoretical claims, experimental protocols, interpretation of results, and conclusions, were made by us.

We take full responsibility for the final content of this work, including text, mathematical claims, experimental results, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work advances fundamental methodology for generative modeling and does not involve human subjects, sensitive personal data, or deployment in high-stakes applications. We do not identify any specific societal or ethical concerns arising from the contributions of this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. For the theoretical results, the Appendix provides the full mathematical settings, assumptions, statements, and proofs. For the empirical results, the Appendix provides detailed descriptions of the datasets and benchmarks, experimental protocols, model and training configurations, evaluation procedures, and relevant hyperparameters. Upon acceptance, we will release the source code required to reproduce our experiments in a public GitHub repository and provide the corresponding repository link in the paper.

REFERENCES

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. *Advances in Neural Information Processing Systems*, 38:8522–8549, 2026.

Samuel Boïté, Julie Delon, and Kimia Nadjahi. Expected batch optimal transport plans and consequences for flow matching. *arXiv preprint arXiv:2605.12174*, 2026.

Rania Briq, Michael Kamp, Ohad Fried, Sarel Cohen, and Stefan Kesselheim. The amazing stability of flow matching. *arXiv preprint arXiv:2604.16079*, 2026.

Chiensheng Chiang, Kuan-Hsun Tu, Jia-Wei Liao, Cheng-Fu Chou, and Tsung-Wei Ke. COT-FM: Cluster-wise optimal transport flow matching. *arXiv preprint arXiv:2603.13395*, 2026.

David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (OTT): A JAX toolbox for all things Wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-First International Conference on Machine Learning*, 2024.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Ka-Hei Hui, Chao Liu, Chi-Wing Fu, Arash Vahdat, et al. Not-so-optimal transport flows for 3d point cloud generation. In *International Conference on Learning Representations*, volume 2025, pp. 98748–98771, 2025.

Beomsu Kim, Yu-Guan Hsieh, Michal Klein, Jong Chul Ye, Bahjat Kawar, James Thornton, et al. Simple reflow: Improved techniques for fast flow models. In *International Conference on Learning Representations*, volume 2025, pp. 67740–67766, 2025.

Lingkai Kong, Molei Tao, Yang Liu, Bryan Wang, Jinmiao Fu, Chien-Chih Wang, and Huidong Liu. AlignFlow: Improving flow-based generative models with semi-discrete optimal transport. In *International Conference on Learning Representations*, volume 2026, pp. 132826–132844, 2026.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. *Advances in Neural Information Processing Systems*, 37:104180–104204, 2024.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in Neural Information Processing Systems*, 37:63082–63109, 2024.

Ziyun Li, Huancheng Hu, Soon Hoe Lim, Xuyu Li, Fei Gao, Enmao Diao, Zezhen Ding, Michalis Vazirgiannis, and Henrik Bostrom. A kinetic energy perspective of flow matching. *arXiv preprint arXiv:2602.07928*, 2026.

Yexiong Lin, Yu Yao, Yang Zhou, and Tongliang Liu. Beyond optimal transport: Model-aligned coupling for flow matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3955–3964, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. Fast optimal transport through sliced generalized Wasserstein geodesics. *Advances in Neural Information Processing Systems*, 36:35350–35385, 2023.

Alireza Mousavi-Hosseini, Stephen Zhang, Michal Klein, et al. Flow matching with semidiscrete couplings. In *International Conference on Learning Representations*, volume 2026, pp. 95449–95486, 2026.

Gabriel Peyré. Optimal transport for machine learners. *arXiv preprint arXiv:2505.06589*, 2025.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Emile Pierret, Johannes Hertrich, Samuel Hurault, and Julie Delon. Tessellations of semi-discrete flow matching. *arXiv preprint arXiv:2605.07513*, 2026.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023a.

Aram-Alexandre Pooladian, Vincent Divol, and Jonathan Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning*, pp. 28128–28150. PMLR, 2023b.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. *Advances in Neural Information Processing Systems*, 38:146459–146512, 2026.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Panos Tsimpos and Youssef Marzouk. An Eulerian perspective on straight-line sampling. *arXiv preprint arXiv:2510.11657*, 2025.

Panos Tsimpos, Daniel Sharp, and Youssef Marzouk. One-shot generative flows: Existence and obstructions. *arXiv preprint arXiv:2604.15439*, 2026.

Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2021.

Stephen Zhang, Alireza Mousavi-Hosseini, Michal Klein, and Marco Cuturi. On fitting flow models with large sinkhorn couplings. *arXiv preprint arXiv:2506.05526*, 2025.

APPENDIX

A RELATED WORK

Flow matching (FM) and OT-guided flows. OT (Peyré & Cuturi, 2019) defines a natural geometry on the space of probability measures and provides principled guidance to train flow-based models: OT-CFM and Multisample Flow Matching use minibatch OT, matching batches of noise samples to target samples, to reduce conflicting velocities and obtain straighter flows, improving sampling efficiency at low NFE (Tong et al., 2023; Pooladian et al., 2023a). However, they are not widely used in practice since the batch size needs to be large enough for them to perform well: the optimal matchings on small sample sizes are unstable and cannot reliably approximate matches (Zhang et al., 2025). Kornilov et al. (2024); Korotin et al. (2022); Chiang et al. (2026) exploit OT directly for few-step generation. These works motivate OT as an important ingredient in efficient generation. On the theory side, Tsimpos & Marzouk (2025); Tsimpos et al. (2026) study characterizations, existence, and obstructions of straight flows in flow-based models.

OT-guided flows versus neural realization. Despite the ideal nature of OT, optimal couplings need not be easy for a neural network to realize. Hui et al. (2025) observes that straighter OT supervision can induce a more complex velocity field and introduces hybrid couplings to improve generative performance. Model-Aligned Coupling (MAC) instead asks which couplings the current model can learn most easily and selects pairs according to its prediction error (Lin et al., 2026). Similar to Reflow, MAC may become unreliable for post-training if the pretrained model is poor, since its coupling criterion is itself determined by the model's predictions. These results distinguish transport optimality from neural approximation, which we leverage to extend ReCouple to Recouple-AS and ReCouple++ by additionally controlling spatial regularity of the learned velocity field.

Approximation of OT. Large-batch Sinkhorn methods improve the approximation of OT but increase computational cost, while recent theory characterizes the population coupling induced by minibatch OT and its interaction with numerical integration (Boïté et al., 2026). Min-SWGG and sliced transport plans provide cheaper approximations through low-dimensional projections (Mahey et al., 2023). A complementary line exploits the semi-discrete setting of generation, where the source distribution is continuous but only finitely many target samples are available: Pooladian et al. (2023b) studies statistical properties, whereas SD-FM and AlignFlow propose efficient scheme to construct SDOT couplings between fresh source samples and the empirical data distribution (Mousavi-Hosseini et al., 2026; Kong et al., 2026). ReCouple is not primarily designed as a more accurate or efficient approximation to OT or SDOT. Instead, it introduces a controllable approximation to understand how coupling optimality affects the learnability and few-step behavior of FM models, complementing the recent line of studies (Li et al., 2026; Bertrand et al., 2026; Briq et al., 2026) that aim to provide a deeper understanding of flow-based models.

Rollout-free versus rollout-based post-training. Rectified Flow uses Reflow to progressively straighten model-induced couplings (Liu et al., 2022) but it requires first rolling out a teacher network before retraining. Moreover, Reflow is constrained by the quality of the pretrained flow model, which needs to be good enough to generate approximately the target distribution, not to mention the large compute spent on generating a large number of data points at each rollout-based step. Another line of work accelerates pretrained diffusion or flow models through model generated trajectories (Lee et al., 2024; Kim et al., 2025) or distillation (Sabour et al., 2026). Our framework does not use pretrained-model rollouts or teacher trajectories. Instead, it provides a rollout-free, controllable framework to study how coupling optimality, neural realizability, and finite-NFE generation interact.

B MECHANISTIC INSIGHTS ON RECOUPLE

Why not just select random blocks? A natural question is whether the blockwise procedure based on quantile sorting in ReCouple equivalent to selecting random blocks, and if not, what are the advantages over the random baseline (RandomBlock&Select). ReCouple and RandomBlock&Select use the same sparsity budget but induce very different sparse transport graphs. For simplicity, assume that B divides N throughout this section. Any balanced block construction with block size B retains

$$\frac{N}{B}B^2 = NB \tag{12}$$

admissible source–target edges, corresponding to edge density B/N. Thus the difference between ReCouple and RandomBlock is not how many edges are retained, but which edges are retained.

Recall that in ReCouple, for a direction θ, each sample is assigned its rank in the ordering induced by the projection $x \mapsto \langle \theta, x \rangle$ after the sorting. ReCouple partitions these projected ranks into contiguous blocks of size B and solves the full assignment problem only within corresponding source and target rank blocks. The projection in ReCouple is only 1D, but after data-dependent direction selection the induced sparse OT graph is very different from a uniformly random sparse graph. In particular, ReCouple searches for an ordering (induced by the best direction selected among the R candidates) that makes the OT problem local and sparse.

For comparison, define RandomBlock by conditioning on a balanced labeled source partition

$$S_1, \dots, S_K, \qquad |S_k| = B, \tag{13}$$

and drawing the labeled target partition

$$T_1, \dots, T_K, \qquad |T_k| = B, \tag{14}$$

uniformly from all balanced labeled target partitions. The corresponding block graph contains the edges

$$E_{\mathcal{B}} = \bigcup_{k=1}^{K} S_k \times T_k. \tag{15}$$

Proposition 3 (ReCouple versus random baseline)**.** *Let*

$$C^\star := \frac{1}{N} \min_{\pi \in S_N} \sum_{i=1}^{N} c_{i,\pi(i)} \tag{16}$$

denote the unrestricted empirical quadratic OT cost. For a balanced block geometry $\mathcal{B}$, define

$$\mathcal{F}_{\mathcal{B}} := \{\pi \in S_N : \pi(S_k) = T_k,\ \forall k\}, \tag{17}$$

and

$$C_{\mathcal{B}} := \frac{1}{N} \min_{\pi \in \mathcal{F}_{\mathcal{B}}} \sum_{i=1}^{N} c_{i,\pi(i)}, \tag{18}$$

where $c_{ij} := \|x_i - y_j\|^2$. Then the following statements hold.

(i) Every balanced block geometry of block size B contains exactly $|E_{\mathcal{B}}| = NB$ admissible edges, and therefore has edge density B/N.

(ii) Let $\pi^\star \in S_N$ be any fixed permutation and let

$$\mathrm{Rec}_B^{\mathrm{rand}} := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{(i, \pi^\star(i)) \in E_{\mathcal{B}}\}. \tag{19}$$

For RandomBlock,

$$\mathbb{E}[\mathrm{Rec}_B^{\mathrm{rand}}] = \frac{B}{N}. \tag{20}$$

If $\pi^\star$ is the unique global OT permutation, then

$$\mathbb{P}(C_B^{\mathrm{rand}} = C^\star) = \frac{(B!)^{N/B}}{N!}. \tag{21}$$

(iii) Fix a direction θ. Let

$$\theta^\top x_{i_1} < \cdots < \theta^\top x_{i_N}, \qquad \theta^\top y_{j_1} < \cdots < \theta^\top y_{j_N}, \tag{22}$$

and define the rank-matching permutation $\sigma_\theta(i_r) = j_r$. Let

$$C^\theta_{\mathrm{rank}} := \frac{1}{N} \sum_{r=1}^{N} \|x_{i_r} - y_{j_r}\|^2 \tag{23}$$

denote the quadratic cost of this rank matching and C^θ_B is the ReCouple blockwise cost induced by θ. Then

$$C^\star \le C^\theta_B \le C^\theta_{\mathrm{rank}}. \tag{24}$$

In particular,

$$C^\theta_1 = C^\theta_{\mathrm{rank}}, \qquad C^\theta_N = C^\star. \tag{25}$$

(iv) Let $(u^\star, v^\star)$ be any optimal dual solution of the unrestricted assignment problem and define the costs

$$r_{ij} := c_{ij} - u^\star_i - v^\star_j \ge 0. \tag{26}$$

Then, for any admissible block geometry $\mathcal{B}$,

$$C_{\mathcal{B}} - C^\star = \frac{1}{N} \min_{\pi \in \mathcal{F}_{\mathcal{B}}} \sum_{i=1}^{N} r_{i,\pi(i)}. \tag{27}$$

Detailed proof is in App. D.5.

Proposition 3 clarifies the advantage of ReCouple. The relevant criterion is not whether a sparse graph preserves every edge of one particular optimal permutation. By equation 27, a sparse graph is sufficient whenever it supports *some* perfect matching having small total reduced cost. OT-edge recall is therefore a useful geometric diagnostic, but it is not the optimization objective itself. ReCouple uses projected geometry to generate structured candidate graphs. In contrast, RandomBlock&Select uses the same edge budget and the same best-of-R selection principle, but its candidate balanced block graphs are generated without reference to the data geometry.

To separate the effect of the proposal law from the effect of best-of-R selection, let

$$p_{\mathrm{proj}}(\varepsilon) := \mathbb{P}\left(\Delta^{\mathrm{proj}}_B \le \varepsilon\right), \qquad p_{\mathrm{RB}}(\varepsilon) := \mathbb{P}\left(\Delta^{\mathrm{RB}}_B \le \varepsilon\right), \tag{28}$$

where $\Delta_B := C_B - C^\star$. For R independent candidates from either proposal family,

$$\mathbb{P}\left(\min_{1 \le r \le R} \Delta^{(r)}_B \le \varepsilon\right) = 1 - \left[1 - p(\varepsilon)\right]^R. \tag{29}$$

Hence the difference between ReCouple and RandomBlock&Select is the lower-tail quality of their respective proposal distributions over sparse graphs. There is no universal result asserting $p_{\mathrm{proj}}(\varepsilon) \ge p_{\mathrm{RB}}(\varepsilon)$. Such an advantage is a property of the data geometry and can be measured empirically.

Figure 3 illustrates this distinction in the selected ReCouple projected-rank coordinates for the case where the source samples are from the standard Gaussian and the target samples are from a 4-mode Gaussian. Both methods expose exactly NB admissible edges. In this example, however, the direction selected by ReCouple makes the unrestricted empirical OT permutation noticeably more local in projected rank, so a larger fraction of its edges falls inside the diagonal $B \times B$ block graph. RandomBlock&Select has the same sparsity and candidate budget but does not exploit this ordering.

When $B = N$, both methods coincide and are simply the global empirical OT. However, when $B < N$, they are not the same. ReCouple can approximate OT much faster and has a higher OT-edge recall at a small B compared to the random baseline.

Table 5 shows that a single random projection is not necessarily better than a random block partition, but the best-of-R selection is much more effective for ReCouple. At $B = 8$, increasing R from 1 to 32 reduces the ReCouple's gap from about 1.99 to 0.53, while RandomBlock&Select only improves from about 1.84 to 1.28. Also, ReCouple's global-OT edge recall rises to about 0.42,

Figure 3: Sparse OT graph comparison at $B = 8$ for ReCouple and RandomBlockSelect.

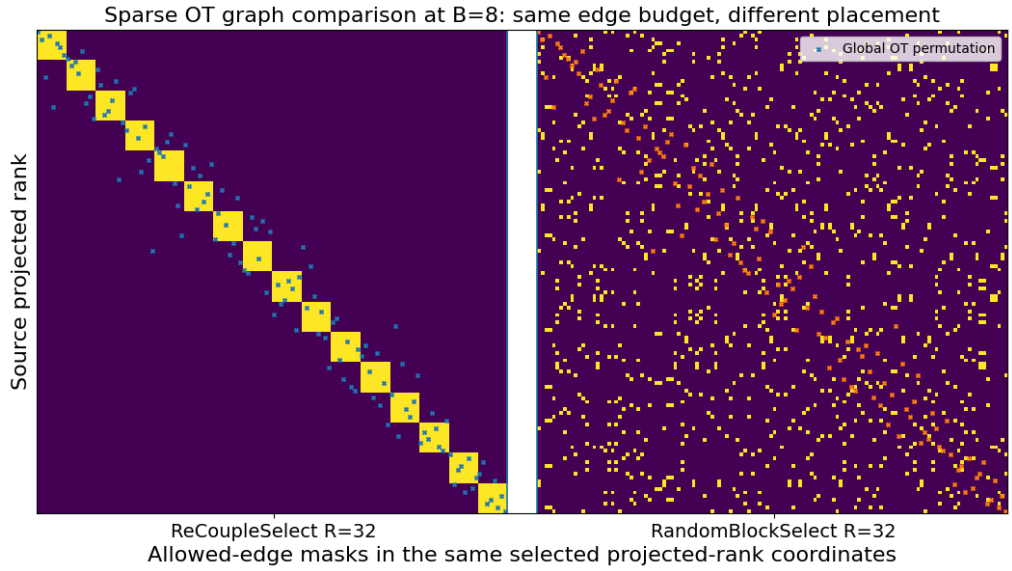


Figure 4: Gap to optimal coupling cost and OT-edge recall across R at $B = 8$. OT-edge recall is the fraction of edges in the global empirical OT matching that are contained in the admissible block graph.

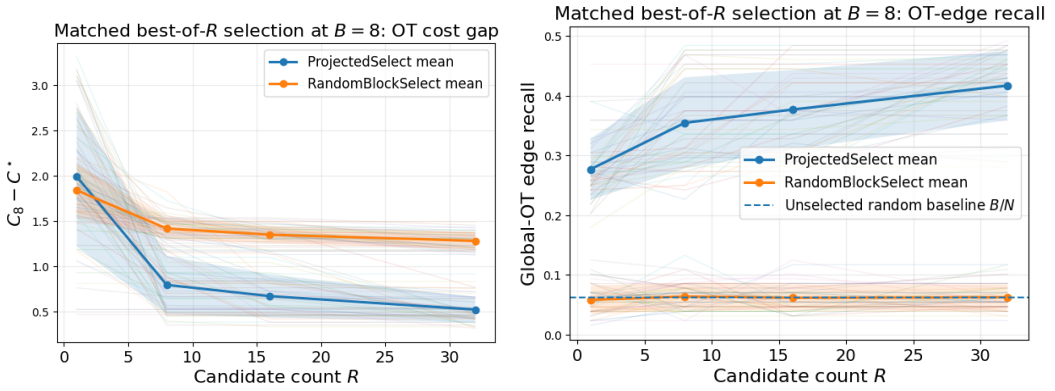


Figure 5: Held-out gap to optimal coupling cost and OT-edge recall across B for various choices of R.

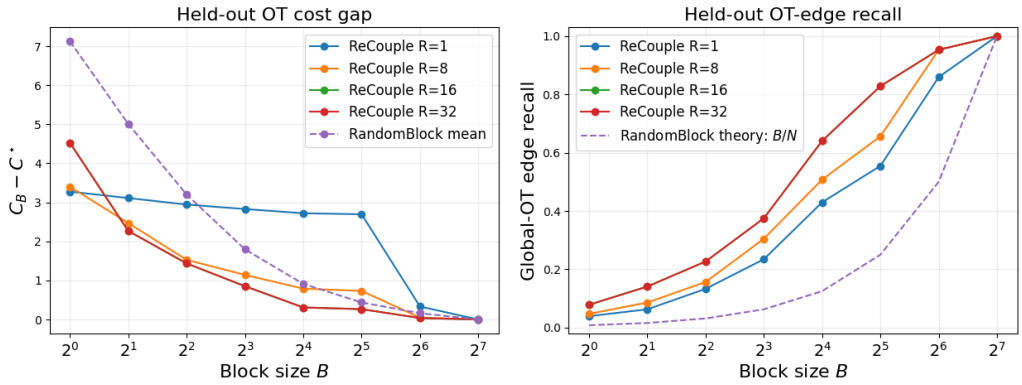


Table 5: Matched best-of-R candidate selection for the standard Gaussian source to four-mode Gaussian target experiment ($N = 128$, $B = 8$). ReCouple chooses the best of R random projection-induced block geometries, while RandomBlock&Select chooses the best of R independent geometry-blind balanced block geometries, using the same block-OT objective. Values are mean $\pm$ standard deviation over 50 independent candidate pools. The unrestricted global empirical OT cost is $C^\star$. OT-edge recall is the fraction of edges of the global OT permutation retained by the selected block graph; the geometry-blind baseline is $B/N = 0.0625$.

	$C_8 - C^\star \downarrow$		OT-edge recall $\uparrow$	
R	ReCouple	RandomBlock&Select	ReCouple	RandomBlock&Select
1	1.991 ± 0.772	$\mathbf{1.840 \pm 0.292}$	$\mathbf{0.277 \pm 0.051}$	0.058 ± 0.024
8	$\mathbf{0.797 \pm 0.319}$	1.418 ± 0.114	$\mathbf{0.355 \pm 0.076}$	0.064 ± 0.023
16	$\mathbf{0.674 \pm 0.236}$	1.351 ± 0.094	$\mathbf{0.377 \pm 0.066}$	0.062 ± 0.022
32	$\mathbf{0.525 \pm 0.141}$	1.283 ± 0.094	$\mathbf{0.417 \pm 0.058}$	0.063 ± 0.022

versus the random baseline $B/N = 0.0625$. Overall these empirical results support the claim that ReCouple's best-of-R search over structured space can find much better sparse OT graph than RandomBlock&Select.

Why use the 1D random slices/projections? ReCouple has the broad structure of "partition, solve OT locally, and combine," but its blocks are not clusters in the usual Euclidean or semantic sense. A unit direction $\theta \in S^{d-1}$ defines the shared scalar score $s_\theta(x) = \langle \theta, x \rangle$ and the source and target samples are partitioned into corresponding equal-size intervals of their projected ranks. The projection is used only to construct the admissible graph.

Other shared scalar scores could in principle be used to construct equal-rank blocks. Linear projections are particularly attractive because they are inexpensive, require no additional metric or bandwidth, and have a direct relation to Euclidean quadratic cost:

$$|\langle \theta, x - y \rangle| \le \|x - y\|, \qquad \mathbb{E}_{\theta \sim \mathrm{Unif}(\mathbb{S}^{d-1})} |\langle \theta, x - y \rangle|^2 = \frac{1}{d} \|x - y\|^2. \tag{30}$$

Finally, ReCouple does not rely on a single projection. It searches over several candidate directions and selects the one having the lowest full-dimensional blockwise OT cost. Its advantage over a purely random block construction is therefore not guaranteed by projection alone; it is more pronounced when the projection proposal family places more probability on graphs supporting low-cost perfect matchings.

Relation with SWGG (Mahey et al., 2023). The core idea of ReCouple is similar in spirit to SWGG: both involves taking the rank matching induced by a random slice/projection and lifts that permutation back to the original d-dimensional samples. However, SWGG uses it to define a distributional discrepancy, whereas ReCouple uses it to define structured feasible sets for OT approximation.

To be precise, let $\widehat{\mu} = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$, $\widehat{\nu} = \frac{1}{N} \sum_{i=1}^N \delta_{z_i}$, and, for $\theta \in \mathbb{S}^{d-1}$, let $\sigma_y^\theta, \sigma_z^\theta \in S_N$ sort the projected samples $\langle \theta, y_i \rangle$ and $\langle \theta, z_i \rangle$. The induced rank coupling is $\pi_\theta\big(\sigma_y^\theta(k)\big) = \sigma_z^\theta(k)$, $k = 1, \dots, N$. For quadratic cost, $C_N(\pi) := \frac{1}{N} \sum_{i=1}^N \|y_i - z_{\pi(i)}\|^2$, the fixed-direction SWGG cost is precisely $\mathrm{SWGG}_2^2(\widehat{\mu}, \widehat{\nu}; \theta) = C_N(\pi_\theta)$, and $\text{min-SWGG}_2^2(\widehat{\mu}, \widehat{\nu}) = \min_{\theta \in \mathbb{S}^{d-1}} C_N(\pi_\theta)$ (Mahey et al., 2023).

For ReCouple, let $\Pi_B(\theta) \subseteq S_N$ denote the permutations that match only within corresponding blocks of B consecutive projected ranks, and define $\pi_{\theta,B} \in \arg\min_{\pi \in \Pi_B(\theta)} C_N(\pi)$. Given directions $\theta_1, \dots, \theta_R$, ReCouple selects $\widehat{\pi}_{B,R} \in \arg\min_{1 \le r \le R} C_N(\pi_{\theta_r,B})$, up to the additional identity candidate used in our implementation.

- When $B = 1$, $\Pi_1(\theta) = \{\pi_\theta\}$, and therefore $\widehat{\pi}_{1,R} \in \arg\min_{1 \le r \le R} \mathrm{SWGG}_2^2(\widehat{\mu}, \widehat{\nu}; \theta_r)$. Thus ReCouple with $B = 1$ is exactly a finite best-of-R search over the projection-induced transport plans underlying min-SWGG (with the identity candidate providing one additional option).

- For $B > 1$, the SWGG rank coupling remains feasible, $\pi_\theta \in \Pi_B(\theta)$, hence $C_N(\pi_{\theta,B}) \leq C_N(\pi_\theta) = \mathrm{SWGG}_2^2(\widehat{\mu}, \widehat{\nu}; \theta)$. Consequently, $W_2^2(\widehat{\mu}, \widehat{\nu}) \leq C_N(\widehat{\pi}_{B,R}) \leq \min_{1 \leq r \leq R} \mathrm{SWGG}_2^2(\widehat{\mu}, \widehat{\nu}; \theta_r)$, where inclusion of the identity candidate can only further decrease the middle quantity. On the other hand, $B = N \implies \Pi_N(\theta) = S_N$, so the inner problem reduces to empirical global OT.

Therefore, ReCouple interpolates between the finite-search min-SWGG plans ($B = 1$) and the global empirical OT ($B = N$).

This connection is to SWGG rather than ordinary sliced Wasserstein. Sliced Wasserstein evaluates the 1D projected transport cost, whereas SWGG lifts the projection-induced rank pairing back to the ambient space and evaluates $\frac{1}{N} \sum_{k=1}^{N} \left\| y_{\sigma_y^\theta(k)} - z_{\sigma_z^\theta(k)} \right\|^2$. ReCouple uses this same projected rank structure to define a feasible set for the original d-dimensional assignment problem.

We end with a caveat. Note that the construction of Mahey et al. (2023) is called min-SW lifted transport plans in (Peyré, 2025) (see Chapter 11.2 there), which exposes that min-SW does not metrize weak convergence. However, the projection is not the transport approximation itself. It is merely a cheap mechanism for constructing a structured sparse support on which one then solves the original ambient assignment problem (see also Figure 11.6 in (Peyré, 2025)). ReCouple is an extension which shares the same principle: it samples R 1D projections to generate candidate coarse rank partitions, solves the exact ambient-space assignment within blocks of size B, and selects the lowest-cost candidate, with B controlling the locality of the transport constraint and R controlling the search over admissible coarse-grainings.

C Detailed Algorithms

In this section, we provide detailed algorithm for the post-training scheme, as well as ReCouple, ReCouple+ and ReCouple++.

Algorithm 2 Vanilla post-training on a fixed coupling

Require: fixed coupling $\widehat{\gamma}$; pretrained backbone v^{w_0} with parameters w_0; EMA rate β_{EMA}; post-training budget J.

1: Initialize $w \leftarrow w_0$ and $\bar{w} \leftarrow w_0$.
2: **for** $j = 1, \ldots, J$ **do**
3: Draw $(Z, Y) \sim \widehat{\gamma}$ and $t \sim \mathrm{Unif}[0, 1]$.
4: Set $X_t = (1 - t)Z + tY$ and $U = Y - Z$.
5: Update w by one optimizer step on $\|v_t^w(X_t) - U\|^2$.
6: Update the EMA parameters $\bar{w} \leftarrow \beta_{\mathrm{EMA}} \bar{w} + (1 - \beta_{\mathrm{EMA}}) w$.
7: **end for**
8: **return** the velocity field $v^{\bar{w}}$.

Algorithm 1 assumes B divides N for simplicity. The following general form allows a possibly smaller final projected-rank block when B does not divide N and reduces to Algorithm 1 when B divides N.

Algorithm 3 ReCouple: general case

Require: target bank $y_{1:N}$; source bank $z_{1:N}$; block size $B \le N$; proposal count R

Add the identity pairing $\pi_0(i) = i$, $i = 1, \dots, N$, to the candidate set.

Set

$$Q = \left\lceil \frac{N}{B} \right\rceil, \qquad I_b = \{(b-1)B + 1, \dots, \min\{bB, N\}\}, \qquad b = 1, \dots, Q.$$

for $r = 1, \dots, R$ **do**

  Draw $\theta_r \sim \mathrm{Unif}(\mathbb{S}^{d-1})$.

  Let $\sigma_r, \tau_r \in S_N$ sort $\{\langle \theta_r, y_i \rangle\}_{i=1}^N$ and $\{\langle \theta_r, z_j \rangle\}_{j=1}^N$, respectively.

  for $b = 1, \dots, Q$ **do**

    Compute

$$\rho_{r,b} \in \arg\min_{\rho \in \mathfrak{S}(I_b)} \sum_{k \in I_b} \left\| y_{\sigma_r(k)} - z_{\tau_r(\rho(k))} \right\|^2.$$

  end for

  Define $\pi_r \in S_N$ by

$$\pi_r(\sigma_r(k)) = \tau_r(\rho_{r,b}(k)), \qquad k \in I_b, \quad b = 1, \dots, Q.$$

  Evaluate

$$C_N(\pi_r) = \frac{1}{N} \sum_{i=1}^N \|y_i - z_{\pi_r(i)}\|^2.$$

end for

Evaluate $C_N(\pi_0) = \frac{1}{N} \sum_{i=1}^N \|y_i - z_{\pi_0(i)}\|^2$.

Select $r^\star \in \arg\min_{r=0,\dots,R} C_N(\pi_r)$, $\widehat{\pi} = \pi_{r^\star}$.

return $\widehat{\gamma} = \frac{1}{N} \sum_{i=1}^N \delta_{(z_{\widehat{\pi}(i)}, y_i)}$.

Algorithm 4 ReCouple+

Require: target bank $y_{1:N}$; source sampler $Z \sim \nu$; designated source bank $z^{(1)}_{1:N}$; block size $B \le N$; proposal count R; pooling factor M.

Run the proposal stage of Algorithm 1 on $(y_{1:N}, z^{(1)}_{1:N})$ and select

$$\theta^\star = \theta_{r^\star}, \qquad r^\star \in \arg\min_{1 \le r \le R} C_N(\pi_r).$$

Form a pooled source bank $z^{(M)}_{1:MN}$ whose first N samples are $z^{(1)}_{1:N}$ and whose remaining samples are independent draws from ν.

Order the target and pooled source banks by projection onto $\theta^\star$, and relabel them in this order.

Partition the ordered target bank into contiguous blocks I_q of size $b_q \le B$ and associate each I_q with the corresponding block $J_q^{(M)}$ of Mb_q pooled sources.

for each block q **do**

  Solve the capacity-M assignment problem (discrete Kantorovich):

$$\min_{A_{ij} \in \{0,1\}} \sum_{i \in I_q} \sum_{j \in J_q^{(M)}} A_{ij} \|y_i - z_j\|^2,$$

  subject to $\sum_{i \in I_q} A_{ij} = 1$ and $\sum_{j \in J_q^{(M)}} A_{ij} = M$.

end for

Concatenate the block solutions to obtain the pooled permutation $\widehat{\pi}^+$ between the MN duplicated target slots and the MN pooled source samples.

return $\widehat{\gamma}^+ = \frac{1}{MN} \sum_{i=1}^N \sum_{m=1}^M \delta_{\left(z^{(M)}_{\widehat{\pi}^+(i,m)}, y_i\right)}$.

Algorithm 5 Post-training on a fixed coupling with annealed spatial Jacobian regularization

Require: fixed coupling $\widehat{\gamma}$; pretrained backbone v^{w_0}; EMA rate β_{EMA}; post-training budget J; tuning parameters $\sigma, \beta_{\mathrm{reg}}, \rho$.

Initialize $w \leftarrow w_0$, $\bar{w} \leftarrow w_0$, and $J_{\mathrm{AS}} = \max\{2, \mathrm{round}(\rho J)\}$.
If $\beta_{\mathrm{reg}} > 0$, on a fixed calibration batch evaluate the initial CFM loss L_0 and AS loss R_0 under w_0, and set $\lambda_0 = \beta_{\mathrm{reg}} L_0 / R_0$; otherwise set $\lambda_0 = 0$.
for $j = 0, \ldots, J-1$ **do**
    Draw $(Z, Y) \sim \widehat{\gamma}$ and $t \sim \mathrm{Unif}[0,1]$.
    Set $X_t = (1-t)Z + tY$ and $U = Y - Z$.
    Set $\mathcal{L}_{\mathrm{CFM}} = \|v_t^w(X_t) - U\|^2$.
    Set $\lambda_j = \frac{\lambda_0}{2}[1 + \cos(\pi j / J_{\mathrm{AS}})]$ if $j < J_{\mathrm{AS}}$, and $\lambda_j = 0$ otherwise.
    if $\lambda_j > 0$ **then**
        Draw $\xi \sim \mathcal{N}(0, I_d)$ and set $\mathcal{R}_\sigma = (2\sigma^2)^{-1}\|v_t^w(X_t + \sigma\xi) - v_t^w(X_t)\|^2$.
    else
        Set $\mathcal{R}_\sigma = 0$.
    end if
    Update w by one optimizer step on $\mathcal{L}_{\mathrm{CFM}} + \lambda_j \mathcal{R}_\sigma$.
    Update $\bar{w} \leftarrow \beta_{\mathrm{EMA}}\bar{w} + (1 - \beta_{\mathrm{EMA}})w$.
end for
return the EMA velocity field $v^{\bar{w}}$.

D THEORETICAL RESULTS AND PROOFS

Recall that the squared 2-Wasserstein distance is defined as:

$$W_2^2(\nu,\mu) := \inf_{\gamma\in\Pi(\nu,\mu)} \int \|x-y\|^2 \,\mathrm{d}\gamma(x,y) = \inf_{\gamma\in\Pi(\nu,\mu)} \mathbb{E}_\gamma \|X-Y\|^2.$$

D.1 PROOF OF PROPOSITION 1

Proof. Since $T = \nabla\phi$ for a convex potential ϕ by Brenier's theorem (Villani, 2021), T is monotone:

$$\langle T(z) - T(z'), z - z'\rangle \geq 0. \tag{31}$$

For $t < 1$, the interpolation map

$$T_t = (1-t)\mathrm{Id} + tT \tag{32}$$

is therefore strongly monotone, since

$$\begin{aligned}\langle T_t(z) - T_t(z'), z - z'\rangle &= (1-t)\|z-z'\|^2 + t\langle T(z) - T(z'), z - z'\rangle \\ &\geq (1-t)\|z-z'\|^2.\end{aligned} \tag{33}$$

Hence T_t is injective for every $t < 1$. Consequently, Z is a measurable function of $X_t = T_t(Z)$, and therefore so is $\dot{X}_t = T(Z) - Z$. It follows that

$$\mathrm{Cov}(\dot{X}_t \mid X_t) = 0 \tag{34}$$

and

$$v_t^*(X_t) = \mathbb{E}[\dot{X}_t \mid X_t] = \dot{X}_t. \tag{35}$$

Since $\dot{X}_t = T(Z) - Z$ is independent of t along each characteristic, its material acceleration vanishes. □

D.2 PROOF OF THEOREM 2 AND COROLLARY 8

The following is a detailed version of Theorem 2.

Theorem 4 (Global-block ReCouple+ converges to SDOT)**.** *Let $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ be fixed, and let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous. Let $(Z_j)_{j\geq 1}$ be i.i.d. with law ν, and define $\widehat{\nu}_{MN} = \frac{1}{MN}\sum_{j=1}^{MN}\delta_{Z_j}$. Consider ReCouple+ with a single global block, $B = N$, and pooling factor M. Let $\widehat{\gamma}^+_{N,M}$ be an optimal coupling produced by the resulting capacitated assignment, and let*

$$\widehat{C}^+_{N,M} = \int_{\mathbb{R}^d\times\mathbb{R}^d} \|z-y\|^2 \, d\widehat{\gamma}^+_{N,M}(z,y). \tag{36}$$

Then there exists an event Ω_0 with $\mathbb{P}(\Omega_0) = 1$ such that, for every $\omega \in \Omega_0$ and for every choice of finite-M optimal couplings $\widehat{\gamma}^+_{N,M}(\omega)$,

$$\widehat{C}^+_{N,M} \longrightarrow W_2^2(\nu,\widehat{\mu}_N) \qquad \text{as } M\to\infty, \tag{37}$$

and

$$W_2\left(\widehat{\gamma}^+_{N,M}, \gamma_N^{\mathrm{SDOT}}\right) \longrightarrow 0, \tag{38}$$

where γ_N^{SDOT} is the unique quadratic optimal coupling between ν and $\widehat{\mu}_N$.

Proof. We proceed in seven steps.

Step 1: For $B = N$, ReCouple+ coincides with empirical OT.

When $B = N$, all N target samples lie in one block. Hence the projection direction and projected ordering impose no restriction: every pooled source sample may be assigned to every target sample.

Replicate each target M times and introduce the labeled target slots

$$y_{i,r} = y_i, \qquad i = 1, \dots, N, \quad r = 1, \dots, M. \tag{39}$$

Each of the MN source samples and each of the MN target slots has mass $1/(MN)$. After scaling by MN, a transport plan between these two equal-weight empirical measures is represented by a doubly stochastic $MN \times MN$ matrix.

By the Birkhoff-von Neumann theorem (see, e.g., Theorem 3.17 in (Peyré, 2025)), every doubly stochastic matrix is a convex combination of permutation matrices. Since the transport objective is linear, an optimum is attained at a permutation matrix. Therefore the optimal transport problem between the pooled source samples and the replicated target slots is exactly the assignment problem solved by global-block ReCouple+.

It remains to relate the replicated target problem to transport toward $\widehat{\mu}_N$. Collapsing the M copies of each y_i maps any feasible coupling for the replicated problem to a coupling in

$$\Pi(\widehat{\nu}_{MN}, \widehat{\mu}_N) \tag{40}$$

without changing its cost. Conversely, any coupling in $\Pi(\widehat{\nu}_{MN}, \widehat{\mu}_N)$ can be lifted to the replicated problem by splitting the total mass assigned to each y_i among its M identical slots; again, the quadratic cost is unchanged. Hence the two transport problems have the same optimal value.

Consequently,

$$\widehat{C}^{+}_{N,M} = W_2^2(\widehat{\nu}_{MN}, \widehat{\mu}_N), \tag{41}$$

and, after collapsing the replicated target slots,

$$\widehat{\gamma}^{+}_{N,M} \in \Pi_{\mathrm{opt}}(\widehat{\nu}_{MN}, \widehat{\mu}_N). \tag{42}$$

Thus, for $B = N$, ReCouple+ solves the finite-source empirical OT problem exactly.

Step 2: Almost-sure W_2 convergence of the empirical source.

Since $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists an event Ω_0 of probability one on which both

$$\widehat{\nu}_n \rightharpoonup \nu \tag{43}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} \|Z_j\|^2 \longrightarrow \int \|z\|^2 \, d\nu(z) \tag{44}$$

hold as $n \to \infty$. We henceforth fix an arbitrary $\omega \in \Omega_0$ and argue deterministically.

Restricting to the subsequence $n = MN$ gives

$$\widehat{\nu}_{MN} \rightharpoonup \nu \tag{45}$$

and

$$\int \|z\|^2 \, d\widehat{\nu}_{MN}(z) \longrightarrow \int \|z\|^2 \, d\nu(z). \tag{46}$$

Weak convergence together with convergence of second moments is equivalent to convergence in W_2, hence

$$W_2(\widehat{\nu}_{MN}, \nu) \longrightarrow 0. \tag{47}$$

Step 3: Convergence of the optimal costs.

Since $\widehat{\mu}_N$ is fixed, the triangle inequality gives

$$\left| W_2(\widehat{\nu}_{MN}, \widehat{\mu}_N) - W_2(\nu, \widehat{\mu}_N) \right| \le W_2(\widehat{\nu}_{MN}, \nu). \tag{48}$$

By equation 47,

$$W_2(\widehat{\nu}_{MN}, \widehat{\mu}_N) \longrightarrow W_2(\nu, \widehat{\mu}_N). \tag{49}$$

Together with equation 41, this yields

$$\widehat{C}^{+}_{N,M} \longrightarrow W_2^2(\nu, \widehat{\mu}_N). \tag{50}$$

Step 4: Tightness of the empirical optimal couplings.

The first and second marginals of $\widehat{\gamma}^{+}_{N,M}$ are respectively $\widehat{\nu}_{MN}$ and $\widehat{\mu}_N$. Therefore

$$\int \left(\|z\|^2 + \|y\|^2\right) d\widehat{\gamma}^{+}_{N,M}(z,y) = \int \|z\|^2 \, d\widehat{\nu}_{MN}(z) + \int \|y\|^2 \, d\widehat{\mu}_N(y). \tag{51}$$

The right-hand side is uniformly bounded in M. Hence the sequence $\{\widehat{\gamma}^{+}_{N,M}\}_{M\geq 1}$ is tight.

By Prokhorov's theorem, every subsequence $\{\widehat{\gamma}^{+}_{N,M_k}\}$ admits a further subsequence $\{\widehat{\gamma}^{+}_{N,M_{k_\ell}}\}$ such that

$$\widehat{\gamma}^{+}_{N,M_{k_\ell}} \rightharpoonup \bar{\gamma} \tag{52}$$

for some probability measure $\bar{\gamma}$ on $\mathbb{R}^d \times \mathbb{R}^d$.

Step 5: Every subsequential limit has the correct marginals.

The coordinate projections are continuous. Hence, taking marginals in equation 52,

$$(\mathrm{proj}_1)_{\#}\bar{\gamma} = \nu, \qquad (\mathrm{proj}_2)_{\#}\bar{\gamma} = \widehat{\mu}_N. \tag{53}$$

Thus

$$\bar{\gamma} \in \Pi(\nu, \widehat{\mu}_N). \tag{54}$$

Step 6: Every subsequential limit is the SDOT coupling.

The quadratic cost

$$c(z,y) = \|z-y\|^2 \tag{55}$$

is nonnegative and lower semicontinuous. Therefore

$$\begin{aligned} \int \|z-y\|^2 \, d\bar{\gamma}(z,y) &\leq \liminf_{\ell\to\infty} \int \|z-y\|^2 \, d\widehat{\gamma}^{+}_{N,M_{k_\ell}}(z,y) \\ &= W_2^2(\nu, \widehat{\mu}_N), \end{aligned} \tag{56}$$

where Step 3 was used in the last equality.

Since $\bar{\gamma} \in \Pi(\nu, \widehat{\mu}_N)$,

$$\int \|z-y\|^2 \, d\bar{\gamma}(z,y) \geq W_2^2(\nu, \widehat{\mu}_N). \tag{57}$$

Hence equality holds, and $\bar{\gamma}$ is an optimal coupling between ν and $\widehat{\mu}_N$.

Because ν is absolutely continuous, Brenier's theorem implies that the quadratic optimal coupling is unique. Therefore

$$\bar{\gamma} = \gamma_N^{\mathrm{SDOT}}. \tag{58}$$

We have shown that every subsequence admits a further subsequence converging weakly to the same limit γ_N^{SDOT}. Consequently,

$$\widehat{\gamma}^{+}_{N,M} \rightharpoonup \gamma_N^{\mathrm{SDOT}} \qquad \text{as } M \to \infty. \tag{59}$$

Step 7: Upgrade weak convergence to W_2 convergence.

With the product Euclidean norm on $\mathbb{R}^d \times \mathbb{R}^d$,

$$\|(z,y)\|^2 = \|z\|^2 + \|y\|^2. \tag{60}$$

Hence

$$\begin{aligned} \int \|(z,y)\|^2 \, d\widehat{\gamma}^{+}_{N,M}(z,y) \; &= \int \|z\|^2 \, d\widehat{\nu}_{MN}(z) + \int \|y\|^2 \, d\widehat{\mu}_N(y) \\ &\longrightarrow \int \|z\|^2 \, d\nu(z) + \int \|y\|^2 \, d\widehat{\mu}_N(y) = \int \|(z,y)\|^2 \, d\gamma_N^{\mathrm{SDOT}}(z,y). \end{aligned} \tag{61}$$

Combining this second-moment convergence with equation 59 yields

$$W_2\left(\widehat{\gamma}^{+}_{N,M}, \gamma_N^{\mathrm{SDOT}}\right) \longrightarrow 0. \tag{62}$$

Since the argument holds for every $\omega \in \Omega_0$ and for every choice of finite-M optimal couplings, the result follows almost surely. □

ReCouple+ converges to SDOT in the joint limit of $B \to N$ and $M \to \infty$.

Corollary 5 (Joint limit of ReCouple+)**.** *Fix N. Under the assumptions of Theorem 2, let (B_k, M_k) satisfy*

$$B_k \in \{1, \dots, N\}, \qquad B_k \to N, \qquad M_k \to \infty. \tag{63}$$

Then there exists an event Ω_0 with $\mathbb{P}(\Omega_0) = 1$ such that, on Ω_0, for every such sequence (B_k, M_k) and every choice of optimal ReCouple+ couplings,

$$\widehat{C}^{+}_{N,B_k,M_k} \longrightarrow W_2^2(\nu, \widehat{\mu}_N), \tag{64}$$

and

$$W_2\Big(\widehat{\gamma}^{+}_{N,B_k,M_k}, \gamma_N^{\mathrm{SDOT}}\Big) \longrightarrow 0. \tag{65}$$

Proof of Corollary 5. Because $B_k \in \{1, \dots, N\}$ and $B_k \to N$, there exists k_0 such that $B_k = N$, for all $k \geq k_0$. Hence, after discarding finitely many terms, ReCouple+ has a single global block and the projected-rank restriction is absent.

By Theorem 2, there exists an event Ω_0 of probability one on which, for every choice of finite-M optimal couplings,

$$\widehat{C}^{+}_{N,N,M} \longrightarrow W_2^2(\nu, \widehat{\mu}_N) \tag{66}$$

and

$$W_2\Big(\widehat{\gamma}^{+}_{N,N,M}, \gamma_N^{\mathrm{SDOT}}\Big) \longrightarrow 0 \tag{67}$$

as $M \to \infty$.

Since $M_k \to \infty$, the same conclusions hold along the subsequence $M = M_k$. The event Ω_0 is independent of the particular sequence (B_k, M_k), so the statement holds simultaneously for every such sequence. □

D.3 The nontrivial limit: fixed $B < N$ and $M \to \infty$.

The more interesting regime keeps the projected-rank restriction fixed while sending the source resolution to infinity. In this case, ReCouple+ converges to a population block-constrained semi-discrete OT problem.

Fix a projection direction $\theta \in \mathbb{S}^{d-1}$ and sort the target samples according to $\langle \theta, y_i \rangle$, with ties broken deterministically. Let

$$I_1, \dots, I_Q \tag{68}$$

denote the resulting consecutive target blocks, with $b_q := |I_q|$ and $\sum_{q=1}^{Q} b_q = N$. Define

$$\alpha_q := \frac{b_q}{N}, \qquad s_q := \sum_{\ell=1}^{q} \alpha_\ell, \qquad s_0 := 0. \tag{69}$$

Let

$$F_\theta(r) := \nu\big(\{z : \langle \theta, z \rangle \leq r\}\big) \tag{70}$$

be the distribution function of the projected source. Since ν is absolutely continuous, its one-dimensional projection is also absolutely continuous, and in particular F_θ is continuous and has no samples. For $q = 1, \dots, Q-1$, let

$$c_q := F_\theta^{-1}(s_q), \tag{71}$$

and set $c_0 = -\infty$ and $c_Q = +\infty$. Define the population rank slabs

$$A_q := \{z : c_{q-1} < \langle \theta, z \rangle \leq c_q\}. \tag{72}$$

Then

$$\nu(A_q) = \alpha_q. \tag{73}$$

Define the normalized source and target block measures

$$\nu_q := \frac{1}{\alpha_q} \nu|_{A_q}, \qquad \mu_q := \frac{1}{b_q} \sum_{i \in I_q} \delta_{y_i}. \tag{74}$$

We first make precise the corresponding population transport problem.

Proposition 6 (Population Block-SDOT). *Let $\Gamma_{B,\theta}$ denote the class of couplings between ν and $\widehat{\mu}_N$ that transport source slab A_q only to targets belonging to block I_q. Then every $\gamma \in \Gamma_{B,\theta}$ admits the decomposition*

$$\gamma = \sum_{q=1}^{Q} \alpha_q \gamma_q, \qquad \gamma_q \in \Pi(\nu_q, \mu_q). \tag{75}$$

Consequently,

$$\inf_{\gamma \in \Gamma_{B,\theta}} \int \|z - y\|^2 \, d\gamma(z, y) = \sum_{q=1}^{Q} \alpha_q W_2^2(\nu_q, \mu_q). \tag{76}$$

Since each ν_q is absolutely continuous, the quadratic optimal coupling $\gamma_q^\star$ between ν_q and μ_q is unique. Hence the constrained problem has the unique optimizer

$$\gamma_{B,\theta}^{\mathrm{BSDOT}} := \sum_{q=1}^{Q} \alpha_q \gamma_q^\star. \tag{77}$$

Proof. The source slabs form a partition of $\mathbb{R}^d$ up to a ν-null set, and the target blocks form a partition of the target samples. The block constraint therefore forces exactly α_q units of mass to be transported from A_q to the targets in I_q. Normalizing the restriction of γ to this block gives a coupling $\gamma_q \in \Pi(\nu_q, \mu_q)$, yielding

$$\gamma = \sum_{q=1}^{Q} \alpha_q \gamma_q. \tag{78}$$

The quadratic cost separates accordingly:

$$\int \|z - y\|^2 \, d\gamma = \sum_{q=1}^{Q} \alpha_q \int \|z - y\|^2 \, d\gamma_q. \tag{79}$$

Minimizing each term independently gives equation 76. Since $\nu_q \ll \mathcal{L}^d$, Brenier's theorem gives uniqueness of each quadratic optimal coupling $\gamma_q^\star$, and hence uniqueness of $\gamma_{B,\theta}^{\mathrm{BSDOT}}$. □

The following lemma is the key statistical step.

Lemma 1 (Convergence of empirical rank blocks). *Let $(Z_j)_{j \geq 1}$ be i.i.d. with law ν. For $n = MN$, sort $Z_1, \ldots, Z_n$ according to $\langle \theta, Z_j \rangle$. For block q, let $\mathcal{I}_{q,M}$ contain the source indices having projected ranks*

$$Ms_{q-1}N + 1, \ldots, Ms_qN, \tag{80}$$

equivalently the Mb_q consecutive source ranks corresponding to target block I_q, and define

$$\widehat{\nu}_{q,M} := \frac{1}{Mb_q} \sum_{j \in \mathcal{I}_{q,M}} \delta_{Z_j}. \tag{81}$$

Then, simultaneously for $q = 1, \ldots, Q$,

$$W_2(\widehat{\nu}_{q,M}, \nu_q) \longrightarrow 0 \qquad \text{a.s. as } M \to \infty. \tag{82}$$

Proof. Write

$$R_j := \langle \theta, Z_j \rangle, \qquad n := MN, \tag{83}$$

and let

$$m_q := \sum_{\ell=1}^{q} b_\ell, \qquad r_{q,M} := Mm_q = ns_q. \tag{84}$$

Because the projected law is atomless, with probability one there are no ties among the R_j.

Let

$$K_{q,M} := \sum_{j=1}^{n} \mathbf{1}_{\{R_j \leq c_q\}}, \tag{85}$$

with $K_{0,M} = 0$ and $K_{Q,M} = n$. By the strong law of large numbers,

$$\frac{K_{q,M}}{n} \longrightarrow F_\theta(c_q) = s_q \qquad \text{a.s.} \tag{86}$$

Hence

$$\frac{|K_{q,M} - r_{q,M}|}{n} \longrightarrow 0 \qquad \text{a.s.} \tag{87}$$

Let

$$\mathcal{J}_{q,M} := \{1 \le j \le n : Z_j \in A_q\} \tag{88}$$

be the indices falling in the fixed population slab. The empirical rank block $\mathcal{I}_{q,M}$ and the population-slab index set $\mathcal{J}_{q,M}$ differ only through discrepancies at their two boundaries. Therefore

$$|\mathcal{I}_{q,M} \triangle \mathcal{J}_{q,M}| \le |K_{q,M} - r_{q,M}| + |K_{q-1,M} - r_{q-1,M}|. \tag{89}$$

By equation 87,

$$\frac{|\mathcal{I}_{q,M} \triangle \mathcal{J}_{q,M}|}{n} \longrightarrow 0. \tag{90}$$

Let f be bounded and continuous. Since $Mb_q = n\alpha_q$,

$$\int f \, d\widehat{\nu}_{q,M} = \frac{1}{n\alpha_q} \sum_{j \in \mathcal{I}_{q,M}} f(Z_j). \tag{91}$$

Using equation 90,

$$\begin{aligned} &\left| \frac{1}{n\alpha_q} \sum_{j \in \mathcal{I}_{q,M}} f(Z_j) - \frac{1}{n\alpha_q} \sum_{j=1}^{n} f(Z_j) \mathbf{1}_{\{Z_j \in A_q\}} \right| \\ &\qquad \le \frac{\|f\|_\infty}{n\alpha_q} |\mathcal{I}_{q,M} \triangle \mathcal{J}_{q,M}| \longrightarrow 0. \end{aligned} \tag{92}$$

The strong law gives

$$\frac{1}{n\alpha_q} \sum_{j=1}^{n} f(Z_j) \mathbf{1}_{\{Z_j \in A_q\}} \longrightarrow \frac{\mathbb{E}[f(Z) \mathbf{1}_{\{Z \in A_q\}}]}{\alpha_q} = \int f \, d\nu_q. \tag{93}$$

Hence

$$\widehat{\nu}_{q,M} \rightharpoonup \nu_q. \tag{94}$$

It remains to prove convergence of second moments. Let

$$H_j := \|Z_j\|^2. \tag{95}$$

The difference between the second moment over the empirical rank block and that over the fixed slab is bounded by

$$\frac{1}{n\alpha_q} \sum_{j \in \mathcal{I}_{q,M} \triangle \mathcal{J}_{q,M}} H_j. \tag{96}$$

For any $K > 0$,

$$\frac{1}{n} \sum_{j \in \mathcal{I}_{q,M} \triangle \mathcal{J}_{q,M}} H_j \le K \, \frac{|\mathcal{I}_{q,M} \triangle \mathcal{J}_{q,M}|}{n} + \frac{1}{n} \sum_{j=1}^{n} H_j \mathbf{1}_{\{H_j > K\}}. \tag{97}$$

The first term converges to zero by equation 90. By the strong law, the limsup of the second term equals

$$\mathbb{E}[\|Z\|^2 \mathbf{1}_{\{\|Z\|^2 > K\}}]. \tag{98}$$

Since $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, this quantity tends to zero as $K \to \infty$. Therefore equation 96 converges to zero.

A final application of the strong law of large number gives

$$\frac{1}{n\alpha_q} \sum_{j=1}^{n} \|Z_j\|^2 \mathbf{1}_{\{Z_j \in A_q\}} \longrightarrow \int \|z\|^2 \, d\nu_q(z). \tag{99}$$

Thus

$$\int \|z\|^2 \, d\widehat{\nu}_{q,M}(z) \longrightarrow \int \|z\|^2 \, d\nu_q(z). \tag{100}$$

Together with equation 94, this is equivalent to equation 82. □

We can now state the fixed-block limit.

Theorem 7 (Fixed-block ReCouple+ converges to Block-SDOT)**.** *Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous, and fix $\widehat{\mu}_N$, θ, and the target blocks $I_1, \dots, I_Q$ defined above. Construct ReCouple+ from the nested source sequence $(Z_j)_{j\geq 1}$ by sorting the MN pooled source samples according to $\langle \theta, Z_j \rangle$, assigning the Mb_q corresponding consecutive source ranks to target block I_q, and solving the capacitated assignment problem within each block.*

Let $\widehat{\gamma}^{+,\theta}_{N,B,M}$ denote any resulting optimal block-constrained coupling, and let $\widehat{C}^{+,\theta}_{N,B,M}$ denote its quadratic cost. Then there exists an event Ω_0 with $\mathbb{P}(\Omega_0) = 1$ such that, on Ω_0, for every choice of finite-M blockwise optimal couplings,

$$W_2\left(\widehat{\gamma}^{+,\theta}_{N,B,M}, \gamma^{\mathrm{BSDOT}}_{B,\theta}\right) \longrightarrow 0 \qquad \text{as } M \to \infty, \tag{101}$$

and

$$\widehat{C}^{+,\theta}_{N,B,M} \longrightarrow \sum_{q=1}^{Q} \alpha_q W_2^2(\nu_q, \mu_q). \tag{102}$$

Proof. Fix an outcome in the probability-one event on which Lemma 1 holds simultaneously for all $q = 1, \dots, Q$. We argue deterministically on this event.

Step 1: Each finite block problem is exact empirical OT.

For block q, the empirical source measure $\widehat{\nu}_{q,M}$ contains Mb_q equally weighted source samples. Each target y_i, $i \in I_q$, has capacity M. Replicating every target in I_q exactly M times converts the capacitated problem into an equal-weight assignment between Mb_q source samples and Mb_q target slots.

By the Birkhoff–von Neumann theorem, the minimum of the corresponding transport linear program is attained at a permutation. Therefore the blockwise assignment solved by ReCouple+ is exactly the quadratic optimal transport problem

$$\widehat{\gamma}_{q,M} \in \Pi_{\mathrm{opt}}(\widehat{\nu}_{q,M}, \mu_q), \tag{103}$$

and its normalized block cost is

$$C_{q,M} = W_2^2(\widehat{\nu}_{q,M}, \mu_q). \tag{104}$$

The global empirical coupling decomposes as

$$\widehat{\gamma}^{+,\theta}_{N,B,M} = \sum_{q=1}^{Q} \alpha_q \widehat{\gamma}_{q,M}, \tag{105}$$

and therefore

$$\widehat{C}^{+,\theta}_{N,B,M} = \sum_{q=1}^{Q} \alpha_q C_{q,M}. \tag{106}$$

Step 2: Convergence of each block cost.

By Lemma 1,

$$W_2(\widehat{\nu}_{q,M}, \nu_q) \longrightarrow 0. \tag{107}$$

Since μ_q is fixed, the triangle inequality gives

$$|W_2(\widehat{\nu}_{q,M}, \mu_q) - W_2(\nu_q, \mu_q)| \leq W_2(\widehat{\nu}_{q,M}, \nu_q) \longrightarrow 0. \tag{108}$$

Hence

$$C_{q,M} \longrightarrow W_2^2(\nu_q, \mu_q). \tag{109}$$

Using equation 106 and the fact that Q is finite,

$$\widehat{C}^{+,\theta}_{N,B,M} \longrightarrow \sum_{q=1}^{Q} \alpha_q W_2^2(\nu_q, \mu_q), \tag{110}$$

which proves equation 102.

Step 3: Convergence of each block coupling.

Fix q. The marginals of $\widehat{\gamma}_{q,M}$ are $\widehat{\nu}_{q,M}$ and μ_q, whose second moments are uniformly bounded. Hence $\{\widehat{\gamma}_{q,M}\}_M$ is tight.

Consider an arbitrary subsequence. By Prokhorov's theorem it admits a further weakly convergent subsequence

$$\widehat{\gamma}_{q,M_\ell} \rightharpoonup \bar{\gamma}_q. \tag{111}$$

The convergence of the marginals implies

$$\bar{\gamma}_q \in \Pi(\nu_q, \mu_q). \tag{112}$$

By lower semicontinuity of the quadratic cost and equation 109,

$$\begin{aligned} \int \|z-y\|^2 \, d\bar{\gamma}_q &\le \liminf_{\ell\to\infty} \int \|z-y\|^2 \, d\widehat{\gamma}_{q,M_\ell} \\ &= W_2^2(\nu_q, \mu_q). \end{aligned} \tag{113}$$

The reverse inequality follows because $\bar{\gamma}_q \in \Pi(\nu_q, \mu_q)$. Hence $\bar{\gamma}_q$ is optimal.

Since ν_q is absolutely continuous, the quadratic optimal coupling between ν_q and μ_q is unique. Therefore

$$\bar{\gamma}_q = \gamma_q^\star. \tag{114}$$

Every subsequence therefore admits a further subsequence converging to $\gamma_q^\star$, and consequently

$$\widehat{\gamma}_{q,M} \rightharpoonup \gamma_q^\star. \tag{115}$$

Moreover,

$$\begin{aligned} \int (\|z\|^2 + \|y\|^2) \, d\widehat{\gamma}_{q,M} &= \int \|z\|^2 \, d\widehat{\nu}_{q,M} + \int \|y\|^2 \, d\mu_q \\ &\longrightarrow \int \|z\|^2 \, d\nu_q + \int \|y\|^2 \, d\mu_q. \end{aligned} \tag{116}$$

Thus the weak convergence upgrades to

$$W_2(\widehat{\gamma}_{q,M}, \gamma_q^\star) \longrightarrow 0 \tag{117}$$

on $\mathbb{R}^d \times \mathbb{R}^d$.

Step 4: Reassemble the blocks.

Using equation 105 and equation 77,

$$\widehat{\gamma}^{+,\theta}_{N,B,M} = \sum_{q=1}^{Q} \alpha_q \widehat{\gamma}_{q,M}, \qquad \gamma^{\mathrm{BSDOT}}_{B,\theta} = \sum_{q=1}^{Q} \alpha_q \gamma_q^\star. \tag{118}$$

For each q, let $\Lambda_{q,M}$ be an optimal W_2 coupling between $\widehat{\gamma}_{q,M}$ and $\gamma_q^\star$. Then

$$\Lambda_M := \sum_{q=1}^{Q} \alpha_q \Lambda_{q,M} \tag{119}$$

is a coupling between $\widehat{\gamma}^{+,\theta}_{N,B,M}$ and $\gamma^{\mathrm{BSDOT}}_{B,\theta}$. Therefore

$$W_2^2\left(\widehat{\gamma}^{+,\theta}_{N,B,M}, \gamma^{\mathrm{BSDOT}}_{B,\theta}\right) \le \sum_{q=1}^{Q} \alpha_q W_2^2(\widehat{\gamma}_{q,M}, \gamma_q^\star) \longrightarrow 0, \tag{120}$$

using equation 117 and the finiteness of Q. This proves equation 101. □

Corollary 8 (Block-SDOT cost gap)**.** *The Block-SDOT cost satisfies*

$$\sum_{q=1}^{Q} \alpha_q W_2^2(\nu_q, \mu_q) \geq W_2^2(\nu, \widehat{\mu}_N). \tag{121}$$

Equality holds if and only if the unique SDOT coupling γ_N^{SDOT} respects the block constraint, in which case

$$\gamma_{B,\theta}^{\mathrm{BSDOT}} = \gamma_N^{\mathrm{SDOT}}. \tag{122}$$

Proof. The admissible Block-SDOT class $\Gamma_{B,\theta}$ is a subset of the full coupling class $\Pi(\nu, \widehat{\mu}_N)$, so restricting the feasible set can only increase the optimal quadratic cost. This proves equation 121. If equality holds, the unique optimizer of the unrestricted problem is also feasible for the constrained problem, hence must equal $\gamma_{B,\theta}^{\mathrm{BSDOT}}$. The converse is immediate. □

D.4 Spatial Regularity for SDOT

Theorem 9 (Spatial regularity for SDOT)**.** *Let $T = T_N^\star$ be the SDOT map from an absolutely continuous source ν to $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$, and let $\{L_i\}_{i=1}^N$ denote its Laguerre cells, so that $T(z) = y_i$ for ν-a.e. $z \in L_i$. Consider the displacement interpolation $X_t = (1-t)Z + tT(Z),\ 0 \leq t < 1$, and let $v_t^\star$ denote its marginal velocity field. For each i, define $L_i^t := (1-t)L_i + ty_i$.*

(i) Cellwise regularity. For $t \in [0,1)$, the marginal velocity admits the representation $\bar{v}_t^\star(x) = \frac{y_i - x}{1-t}$ for $x \in \mathrm{int}(L_i^t)$, which agrees with $v_t^\star$ p_t-a.e., and hence $\nabla_x \bar{v}_t^\star(x) = -\frac{1}{1-t} I_d$, and

$$\|\nabla_x \bar{v}_t^\star(x)\|_{\mathrm{op}} = \frac{1}{1-t}, \qquad \|\nabla_x \bar{v}_t^\star(x)\|_F^2 = \frac{d}{(1-t)^2}. \tag{123}$$

(ii) Global stitching barrier. Suppose that two distinct Laguerre cells L_i and L_j, with $y_i \neq y_j$, have a common boundary point, $\overline{L_i} \cap \overline{L_j} \neq \emptyset$. Fix $t \in (0,1)$, and let $\widetilde{v}_t : \mathbb{R}^d \to \mathbb{R}^d$ be any globally Lipschitz field agreeing with $\bar{v}_t^\star$ on the interiors of the cells. Then $\mathrm{Lip}(\widetilde{v}_t) \geq \max\left\{\frac{1}{t}, \frac{1}{1-t}\right\}$. In particular, if $\widetilde{v}_t \in C^1(\mathbb{R}^d; \mathbb{R}^d)$, then

$$\sup_{x \in \mathbb{R}^d} \|\nabla_x \widetilde{v}_t(x)\|_{\mathrm{op}} \geq \max\left\{\frac{1}{t}, \frac{1}{1-t}\right\}, \quad \sup_{x \in \mathbb{R}^d} \|\nabla_x \widetilde{v}_t(x)\|_F \geq \max\left\{\frac{1}{t}, \frac{\sqrt{d}}{1-t}\right\}. \tag{124}$$

Proof. For part (i), let $z \in \mathrm{int}(L_i)$. Since $T(z) = y_i$,

$$x = (1-t)z + ty_i. \tag{125}$$

For $t < 1$, this can be rewritten as:

$$z = \frac{x - ty_i}{1-t}. \tag{126}$$

The velocity along the corresponding displacement trajectory is $T(z) - z$, and hence

$$T(z) - z = y_i - \frac{x - ty_i}{1-t} = \frac{y_i - x}{1-t}. \tag{127}$$

Since the OT map for quadratic cost is monotone, the interpolation map $S_t = (1-t)I + tT$ is strongly monotone for $t < 1$:

$$\langle S_t(z) - S_t(z'), z - z' \rangle \geq (1-t)\|z - z'\|^2. \tag{128}$$

Thus S_t is injective, up to the usual null-set ambiguity of T, so the transported cell interiors define unambiguous branches of the marginal velocity. Differentiating gives equation 123.

For part (ii), the within-cell identity already implies

$$\mathrm{Lip}(\widetilde{v}_t) \geq \frac{1}{1-t}. \tag{129}$$

Indeed, for any $x, x' \in \mathrm{int}(L_i^t)$,

$$\|\bar{v}_t^\star(x) - \bar{v}_t^\star(x')\| = \frac{1}{1-t}\|x - x'\|. \tag{130}$$

It remains to establish the cross-cell bound. Choose $z_\star \in \overline{L_i} \cap \overline{L_j}$. Approaching $z_\star$ from the two cell interiors produces the corresponding limiting points

$$x_i^\star = (1-t)z_\star + ty_i, \qquad x_j^\star = (1-t)z_\star + ty_j. \tag{131}$$

Their spatial separation is

$$\|x_i^\star - x_j^\star\| = t\|y_i - y_j\|. \tag{132}$$

The limiting cellwise velocities are

$$v_i^\star = y_i - z_\star, \qquad v_j^\star = y_j - z_\star, \tag{133}$$

and hence

$$\|v_i^\star - v_j^\star\| = \|y_i - y_j\|. \tag{134}$$

Since $\widetilde{v}_t$ is Lipschitz, it is continuous; because it agrees with $\bar{v}_t^\star$ on the cell interiors, it therefore takes these limiting values at $x_i^\star$ and $x_j^\star$. Consequently,

$$\mathrm{Lip}(\widetilde{v}_t) \geq \frac{\|\widetilde{v}_t(x_i^\star) - \widetilde{v}_t(x_j^\star)\|}{\|x_i^\star - x_j^\star\|} = \frac{1}{t}. \tag{135}$$

Together with the within-cell bound, this proves the first lower bound.

If $\widetilde{v}_t$ is C^1, then on $\mathbb{R}^d$

$$\mathrm{Lip}(\widetilde{v}_t) = \sup_x \|\nabla_x \widetilde{v}_t(x)\|_{\mathrm{op}} \tag{136}$$

whenever the right-hand side is finite, which gives the second lower bound.

Moreover, $\|A\|_F \geq \|A\|_{\mathrm{op}}$, while part (i) gives the exact within-cell value

$$\|\nabla_x \bar{v}_t^\star(x)\|_F = \frac{\sqrt{d}}{1-t}. \tag{137}$$

This proves equation 124.

Finally, we remark that at $t = 0$ the two limiting spatial points coincide, $x_i^\star = x_j^\star = z_\star$, whereas their limiting velocities differ by

$$v_i^\star - v_j^\star = y_i - y_j \neq 0. \tag{138}$$

Hence no continuous field can agree with both cellwise branches at the common boundary. □

D.5 Proof of Proposition 3

Proof. We prove each statement separately.

Proof of (i). There are $K = \frac{N}{B}$ matched source–target block pairs. Each pair $S_k \times T_k$ contains exactly B^2 admissible edges. Since the block pairs are disjoint,

$$|E_{\mathcal{B}}| = \sum_{k=1}^{K} |S_k \times T_k| = KB^2 = NB. \tag{139}$$

Dividing by the N^2 edges of the complete bipartite graph gives edge density

$$\frac{|E_{\mathcal{B}}|}{N^2} = \frac{B}{N}. \tag{140}$$

Proof of (ii). Condition on the source partition $S_1, \dots, S_K$. Fix a source index i, and let S_k be the unique source block containing i. The target block T_k paired with S_k is a uniformly random subset of B target indices. Therefore, for the fixed target $\pi^\star(i)$,

$$\mathbb{P}\left(\pi^\star(i) \in T_k\right) = \frac{B}{N}. \tag{141}$$

Equivalently,

$$\mathbb{P}\left((i, \pi^\star(i)) \in E_{\mathcal{B}}\right) = \frac{B}{N}. \tag{142}$$

By linearity of expectation,

$$\begin{aligned}\mathbb{E}[\mathrm{Rec}_B^{\mathrm{rand}}] &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\left((i, \pi^\star(i)) \in E_{\mathcal{B}}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{B}{N} = \frac{B}{N},\end{aligned} \tag{143}$$

which gives equation 20. Now suppose that $\pi^\star$ is the unique global OT permutation. Then

$$C_B^{\mathrm{rand}} = C^\star \tag{144}$$

if and only if $\pi^\star$ remains feasible under the random block restriction. Conditional on the source partition, this requires

$$T_k = \pi^\star(S_k), \qquad k = 1, \dots, K. \tag{145}$$

Thus exactly one labeled target partition is compatible with $\pi^\star$. The number of balanced labeled target partitions is

$$\binom{N}{B}\binom{N-B}{B}\cdots\binom{B}{B} = \frac{N!}{(B!)^K}. \tag{146}$$

Since RandomBlock draws uniformly from these partitions,

$$\mathbb{P}(C_B^{\mathrm{rand}} = C^\star) = \frac{1}{N!/(B!)^K} = \frac{(B!)^K}{N!}. \tag{147}$$

Using $K = N/B$ yields equation 21.

Proof of (iii). Set

$$a_r := \theta^\top x_{i_r}, \qquad b_r := \theta^\top y_{j_r}, \tag{148}$$

so that

$$a_1 < \cdots < a_N, \qquad b_1 < \cdots < b_N. \tag{149}$$

We first show that matching a_r to b_r is optimal for the one-dimensional squared cost. Consider any assignment containing an inversion: for some $r < s$ it matches

$$a_r \longleftrightarrow b_\ell, \qquad a_s \longleftrightarrow b_k, \qquad k < \ell. \tag{150}$$

The difference between the cost before and after uncrossing these two matches is

$$\begin{aligned}&\left[(a_r - b_\ell)^2 + (a_s - b_k)^2\right] - \left[(a_r - b_k)^2 + (a_s - b_\ell)^2\right] \\ &\quad = 2(a_s - a_r)(b_\ell - b_k) \geq 0.\end{aligned} \tag{151}$$

Thus removing an inversion never increases the projected quadratic cost. Repeatedly uncrossing inversions yields the monotone rank matching

$$a_r \longleftrightarrow b_r, \qquad r = 1, \dots, N, \tag{152}$$

which is therefore an optimal one-dimensional quadratic coupling. Now partition the two rank orderings into corresponding consecutive blocks of size B. The permutation σ_θ matches each rank r to the same rank r and is therefore feasible for the ReCouple block geometry. Hence

$$C_B^\theta \leq C_{\mathrm{rank}}^\theta. \tag{153}$$

On the other hand, the ReCouple feasible set is a subset of all permutations,

$$\mathcal{F}_B^\theta \subseteq S_N, \tag{154}$$

so restricting the feasible set cannot improve upon unrestricted OT:

$$C^\star \leq C_B^\theta. \tag{155}$$

Together these give equation 24. If $B = 1$, every block contains a single source rank and a single target rank, so σ_θ is the only feasible permutation and

$$C_1^\theta = C_{\text{rank}}^\theta. \tag{156}$$

If $B = N$, there is one global block, every permutation is feasible, and

$$C_N^\theta = C^\star. \tag{157}$$

Proof of (iv). The dual of the unrestricted assignment problem is

$$\max_{u,v} \frac{1}{N} \left(\sum_{i=1}^N u_i + \sum_{j=1}^N v_j \right) \tag{158}$$

subject to

$$u_i + v_j \leq c_{ij}. \tag{159}$$

Let $(u^\star, v^\star)$ be any optimal dual pair and define r_{ij} by equation 26. Dual feasibility gives $r_{ij} \geq 0$. For any permutation $\pi \in S_N$,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N c_{i,\pi(i)} &= \frac{1}{N} \sum_{i=1}^N \left(u_i^\star + v_{\pi(i)}^\star + r_{i,\pi(i)} \right) \\ &= \frac{1}{N} \left(\sum_{i=1}^N u_i^\star + \sum_{j=1}^N v_j^\star \right) + \frac{1}{N} \sum_{i=1}^N r_{i,\pi(i)}, \end{aligned} \tag{160}$$

where the second equality uses the fact that π is a permutation, so

$$\sum_{i=1}^N v_{\pi(i)}^\star = \sum_{j=1}^N v_j^\star. \tag{161}$$

By strong duality,

$$\frac{1}{N} \left(\sum_i u_i^\star + \sum_j v_j^\star \right) = C^\star. \tag{162}$$

Therefore, for every permutation,

$$\frac{1}{N} \sum_i c_{i,\pi(i)} - C^\star = \frac{1}{N} \sum_i r_{i,\pi(i)}. \tag{163}$$

Minimizing this identity over $\pi \in \mathcal{F}_\mathcal{B}$ gives

$$C_\mathcal{B} - C^\star = \frac{1}{N} \min_{\pi \in \mathcal{F}_\mathcal{B}} \sum_i r_{i,\pi(i)}, \tag{164}$$

which proves equation 27. □

E Details on the Empirical Results and Ablation Studies

E.1 2D Multimodal Task

Data and pretrained model. We use a two-dimensional target consisting of 25 isotropic Gaussian modes arranged on a 5×5 grid, with spacing 0.5 and standard deviation 0.05, rotated by 45° and normalized following the dataset generator. We use $50{,}000$ training samples and an independent $50{,}000$-sample test set, with Gaussian source $\nu = \mathcal{N}(0, I_2)$. The velocity field is a Fourier-feature MLP with four hidden layers of width 256 and SiLU activations, using 64-dimensional spatial and time embeddings. The base model is pretrained with independent CFM, and all post-training methods start from the same validated EMA checkpoint.

ReCouple family. ReCouple uses a persistent Gaussian source bank, block size $B = 512$, and $R = 16$ random projection proposals. For each proposal, source and target samples are sorted by their projected coordinates, divided into corresponding blocks, and matched by the assignment in the original data space. The candidate with the smallest full-bank quadratic cost is retained. We vary $B \in \{512, 1024, 2048\}$ in the block-size ablation below. ReCouple+ uses the same projected geometry with source multiplicity $M = 2$, assigning MN Gaussian source samples to N targets with capacity M per target. The projection selected at $M = 1$ is retained to isolate the effect of source populationization.

ReCouple++ uses the same pooled coupling as ReCouple+ and adds the spatial consistency penalty

$$\mathcal{R}_\sigma = \frac{1}{2\sigma^2}\mathbb{E}_\epsilon \left\| v_\theta(x_t + \sigma\epsilon, t) - v_\theta(x_t, t) \right\|^2, \qquad \epsilon \sim \mathcal{N}(0, I_2).$$

We set σ to 5% of the data scale, calibrate its initial contribution to 25% of the CFM loss, and cosine-anneal its weight to zero over the first 70% of post-training.

Baselines. We compare against the following matched baselines. In Figure 1, *Pretrained* denotes the independently pretrained CFM model without post-training. *PersistentIndependent-FT* continues training on one fixed independent Gaussian–target pairing, thereby controlling for additional optimization on a persistent pair bank. *OT-CFM-FT* (Tong et al., 2023) instead constructs fresh exact minibatch OT couplings at every post-training update. We also evaluate *AlignFlow* (Kong et al., 2026), which uses its coupling-alignment construction with the same pretrained backbone and post-training budget. *SDOT* (Mousavi-Hosseini et al., 2026) uses the global semi-discrete quadratic OT coupling from the continuous Gaussian source to the empirical target, providing the unrestricted OT reference. Finally, *Noise-ReCouple* (cf. Hui et al., 2025) keeps the selected ReCouple pairing but perturbs its source endpoints according to

$$z_\beta = \sqrt{1-\beta}\, z + \sqrt{\beta}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_2),$$

with $\beta = 0.01$, providing a control for improving learnability by source smoothing rather than spatial regularization. Whenever applicable, baselines use the same network architecture, pretrained initialization, optimizer, update budget, EMA, and evaluation protocol.

Baseline implementation details. OT-CFM-FT uses fresh exact minibatch OT couplings of size 128 at every update. Noise-ReCouple uses the same persistent ReCouple pairs with source perturbation parameter $\beta = 0.01$. AlignFlow follows the published 2D coupling construction, using an entropic dual fit and a balanced fixed Gaussian pair bank, followed by the common post-training protocol. For SDOT, we solve the unrestricted semi-discrete quadratic OT problem from $\mathcal{N}(0, I_2)$ to the full empirical target using OTT-JAX (Cuturi et al., 2022); during post-training, fresh Gaussian samples are mapped through the resulting fixed Laguerre map at every update. Thus the reported SDOT results correspond to the post-trained neural network rather than direct evaluation of the SDOT map.

Given a pretrained FM backbone, teacher Reflow and distillation method use 500k teacher-generated pairs and 50k student updates. Distillation uses endpoint MSE without LPIPS or a preceding reflow stage. As for Reflow, note that the training can be improved further by using a stronger backbone, adding task-specific training tricks and increase compute budgets (see, e.g., (Lee et al., 2024)) to narrow the gap to state-of-the-art distillation schemes. However, doing so would introduce several additional confounding factors, making it much harder to isolate the effect of the mechanism studied here.

Post-training and evaluation. All fine-tuned models use $10{,}000$ updates, batch size 2048, AdamW with learning rate 10^{-4} and weight decay 10^{-5}, gradient clipping at 10, and EMA decay 0.999. Generation uses fresh Gaussian initial conditions and uniform Euler integration at $\mathrm{NFE} \in \{1, 2, 4, 8\}$. We generate $20{,}000$ samples per method. Empirical W_2 is computed by exact Hungarian matching on 1000-point real/generated subsamples and averaged over three fixed subsampling seeds.

We additionally report the neural material-acceleration energy

$$\mathcal{E}_{\mathrm{acc}} = \int_0^1 \mathbb{E}\,\|\partial_t v_t(X_t) + \nabla_x v_t(X_t) v_t(X_t)\|^2\, dt,$$

estimated by automatic differentiation along 512 common Gaussian trajectories using 100 uniform Euler steps.

Coupling resolution versus SDOT. Table 6 compares the endpoint transport objective and construction cost of pooled ReCouple with global SDOT. SDOT attains a slightly smaller quadratic transport cost in this experiment, whereas ReCouple+ is substantially cheaper to construct. This comparison illustrates the intended role of (B, M): ReCouple+ provides a controlled intermediate approximation rather than an attempt to outperform a dedicated SDOT solver in transport cost.

Table 6: Coupling-construction diagnostics on the 2D task. Preparation times exclude neural post-training.

Coupling	Quadratic cost ↓	Source/target representation	Preparation time (s) ↓
ReCouple+ ($B = 512, M = 2$)	0.7234	100k source / 50k target	159.55
Global SDOT	**0.6692**	continuous source / 50k target	815.19

Additional distributional diagnostics. Tables 7–8 show that the low W_2 values of the ReCouple family are accompanied by strong mode coverage rather than concentration on a subset of modes. At NFE 1, ReCouple and its variants substantially improve recall and coverage over the pretrained and persistent-independent models, with ReCouple achieving coverage 0.7357 and ReCouple+ the smallest sliced W_2. By NFE 8, all ReCouple variants attain recall above 0.99 and coverage around 0.90.

The comparison with SDOT is also informative. SDOT attains the strongest precision/recall-based scores at NFE 8, while the ReCouple family obtains substantially smaller sliced W_2. Thus improved endpoint transport cost or trajectory straightness does not induce a uniform ordering across distributional metrics after neural realization. This is consistent with our motivation for treating coupling resolution and spatial realizability as distinct design considerations.

Table 7: Additional distributional metrics on the 2D task at NFE 1. Higher is better for precision, recall, density, and coverage; lower is better for sliced Wasserstein distance (sW2).

Method	Precision ↑	Recall ↑	Density ↑	Coverage ↑	sW2 ↓
Pretrained	0.9934	0.1138	0.9979	0.0423	0.4293
PersistentIndependent-FT	0.9824	0.2443	0.9843	0.0410	0.4163
AlignFlow	0.5872	0.1893	0.4590	0.0449	0.3654
OT-CFM-FT	0.5789	0.8830	0.3972	0.5273	0.0384
SDOT	0.8700	**0.9559**	0.7973	0.7264	0.0373
Noise-ReCouple	0.7038	0.9459	0.5610	0.6626	0.0292
ReCouple	0.8107	0.9477	0.7124	**0.7357**	0.0236
ReCouple-AS	0.8077	0.9498	0.7021	0.7221	0.0236
ReCouple+	0.8191	0.9467	0.7196	0.7318	**0.0233**
ReCouple++	0.8052	0.9493	0.7019	0.7246	0.0235

Table 8: Additional distributional metrics on the 2D task at NFE 8. Higher is better for precision, recall, density, and coverage; lower is better for sliced Wasserstein distance (sW2).

Method	Precision ↑	Recall ↑	Density ↑	Coverage ↑	sW2 ↓
Pretrained	0.5622	0.8803	0.3928	0.5298	0.0732
PersistentIndependent-FT	0.5492	0.8315	0.3796	0.5073	0.0748
AlignFlow	0.6503	0.9343	0.5044	0.6226	0.0609
OT-CFM-FT	0.9125	0.9781	0.8713	0.8437	0.0176
SDOT	**0.9332**	**0.9948**	**0.8901**	**0.9137**	0.0381
Noise-ReCouple	0.9131	0.9920	0.8650	0.9013	0.0171
ReCouple	0.9110	0.9912	0.8677	0.8999	0.0172
ReCouple-AS	0.9072	0.9945	0.8589	0.9016	0.0173
ReCouple+	0.9122	0.9917	0.8700	0.9058	0.0173
ReCouple++	0.9056	0.9915	0.8576	0.9046	0.0173

E.2 CIFAR-10 Task

Data and pretrained model. We consider unconditional generation of CIFAR-10 images at 32×32 resolution, using the 50,000 training images normalized to $[-1, 1]$. We use two pretrained backbones: rectified flow (RF), trained with independent couplings, and OT-CFM, trained with exact minibatch OT. Both use the same U-Net with 128 base channels, channel multipliers $(1, 2, 2, 2)$, four residual blocks per resolution, and dropout 0.15, pretrained for 234,375 updates with seed 10. Post-training starts from the corresponding EMA weights. ReCouple uses the fixed, unflipped training set and standard Gaussian source samples.

ReCouple family. The default ReCouple configuration uses block size $B = 2048$ and $R = 32$ projection candidates. Each candidate sorts sources and targets along a projection and solves exact quadratic-cost assignments within rank blocks; selection minimizes the full pixel-space coupling cost, including an independent fallback. ReCouple-AS adds annealed spatial regularization with initial penalty-to-CFM loss ratio $\eta = 0.25$. ReCouple+ uses one pooled capacitated transport plan with $M = 10$ Gaussian sources per target, and ReCouple++ adds the same regularization with $\eta = 0.5$. Perturbation magnitude is 0.05 times the RMS per-coordinate data standard deviation. The penalty coefficient is calibrated on 128 pairs and cosine-annealed to zero over 35,000 updates, followed by 15,000 pure-CFM updates. We also include global exact OT ($B = N$) and Noise-ReCouple, which uses $B = 1024$, $R = 8$, and replaces each paired source by $\sqrt{0.99}z + \sqrt{0.01}\epsilon$.

Baselines. We compare against the unchanged pretrained backbone, fresh-independent fine-tuning, exact minibatch OT fine-tuning (OT-CFM-FT) (Tong et al., 2023), AlignFlow (Kong et al., 2026), and SD-FM (Mousavi-Hosseini et al., 2026). AlignFlow and SD-FM use semidiscrete OT; AlignFlow includes horizontally flipped targets, whereas SD-FM uses the same 50,000 unflipped targets as ReCouple. Fresh-independent and OT-CFM fine-tuning use random horizontal flips. Teacher reflow and one-step distillation share 500,000 Gaussian–endpoint pairs generated by accurately integrating the frozen pretrained teacher, using uniform-time CFM's MSE and endpoint MSE at $t = 0$, respectively.

Post-training and evaluation. Each method receives 50,000 updates with batch size 128, a fresh Adam optimizer, learning rate 2×10^{-4}, 5,000-step warmup, gradient clipping at 1, and EMA decay 0.9999. Training seeds are $\{10, 20, 30\}$, with fixed pretrained backbones and shared prepared couplings. We evaluate EMA models using uniform Euler integration at NFE $\{1, 2, 4, 8, 10, 50\}$ and compute FID from 50,000 generated images against CIFAR-10 training statistics using CleanFID's `legacy_tensorflow` mode. Fine-tuned models share evaluation seed 9901; results report mean and sample standard deviation across training seeds (see Table 10-11). Backbone repetitions instead vary the sampling seed. Distillation results beyond NFE 1 are diagnostic. Preparation and post-training times are reported separately, excluding pretraining and FID evaluation.

Wall-clock cost. Table 9 reports wall-clock costs for coupling preparation and post-training on an NVIDIA GH200. Ordinary ReCouple requires 30.53 minutes of one-time coupling construction,

compared with 542.46 minutes for the global $B = N$ assignment, a roughly $17.8\times$ reduction in preparation time in our implementation. Its post-training cost remains comparable to the other fine-tuning baselines, at approximately 105 minutes for either pretrained backbone. Increasing the source resolution through ReCouple+ primarily increases the one-time coupling cost: the extreme $M = 10$ pooled construction requires 477.49 minutes, while its post-training time remains close to that of ordinary ReCouple. In contrast, spatial regularization leaves the coupling unchanged but increases optimization time, as ReCouple-AS and ReCouple++ evaluate the additional consistency objective during post-training. These timings are implementation- and hardware-dependent and are intended to characterize the practical cost of the methods rather than their asymptotic complexity.

Table 9: Wall-clock cost comparison for the CIFAR-10 experiments. Times are reported in minutes. Preparation is the one-time method-specific cost for constructing the required coupling, pair bank, or teacher targets and is shared across compatible runs. Post-training time is reported as mean $\pm$ sample standard deviation over three training seeds. FID evaluation, backbone pretraining, and scheduler waiting are excluded. †Teacher preparation uses the recorded historical allocation wall time with four concurrent GPUs and is therefore not directly hardware-normalized against the other preparation times.

	RF backbone		OT-CFM backbone	
Method	Prep.	Post-train	Prep.	Post-train
Teacher Reflow (MSE)	$445.53^{\dagger}$	105.89 ± 0.61	$284.98^{\dagger}$	108.22 ± 3.15
One-step distillation (MSE)	$445.53^{\dagger}$	107.41 ± 3.63	$284.98^{\dagger}$	105.41 ± 2.73
Fresh-independent FT	0.00	96.64 ± 3.30	0.00	96.10 ± 3.27
OT-CFM-FT (Tong et al. (2023))	0.00	97.73 ± 2.83	0.00	97.88 ± 1.86
AlignFlow (Kong et al. (2026))	30.41	118.53 ± 6.19	30.41	115.91 ± 4.02
SD-FM (Mousavi-Hosseini et al. (2026))	38.16	111.40 ± 1.19	38.16	111.62 ± 3.12
Noise-ReCouple (cf. Hui et al. (2025))	2.69	108.18 ± 2.28	2.69	107.17 ± 3.41
ReCouple ($B = N$)	542.46	96.57 ± 2.77	542.46	98.52 ± 2.90
ReCouple	30.53	105.56 ± 1.09	30.53	104.68 ± 3.01
ReCouple-AS	30.53	157.12 ± 3.19	30.53	155.86 ± 3.70
ReCouple+ ($M = 10$)	477.49	105.90 ± 3.25	477.49	105.95 ± 2.99
ReCouple++ ($M = 10$)	477.49	171.85 ± 2.58	477.49	165.69 ± 1.46

E.3 Ablation Studies

Tables 12–13 report generation quality across numerical resolutions, while Table 14 summarizes the corresponding coupling-side diagnostics.

Block size and projection search. Increasing B systematically improves the endpoint transport objective: the normalized gap to global empirical OT decreases from 0.256 at $B = 512$ to 0.211 at $B = 1024$ and 0.169 at $B = 2048$. The downstream benefit is concentrated at very low NFEs: NFE-1 FID decreases from 137.42 ± 1.04 to 130.07 ± 0.26 and 123.16 ± 0.28, and NFE-2 FID from 48.89 ± 0.36 to 47.02 ± 0.05 and 45.72 ± 0.19. From NFE 4 onward, however, $B = 1024$ and $B = 2048$ perform similarly. Thus, reducing the endpoint-cost gap to empirical OT does not translate uniformly into improved generation at every numerical resolution.

Increasing R from 4 to 16 or 32 has no effect for any tested B: the assignment hashes are identical, and hence so are the coupling diagnostics and downstream FIDs. Since the proposal sets are nested, the selected candidate is already present among the first four projections. Thus best-of-R search has saturated at $R = 4$ for these endpoint banks, although this need not hold in general.

Source populationization. The pooling factor M produces a pronounced NFE-dependent crossover. At fixed $B = 1024$ and $R = 16$, the unregularized family is best at $M = 1$ for NFE 1–2, $M = 2$ for NFE 4, $M = 5$ for NFE 8, and $M = 10$ for NFE 10–50. For example, increasing M from 2 to 10 worsens NFE-1 FID from 155.74 ± 0.32 to 195.09 ± 0.73, while improving NFE-50 FID from 8.28 ± 0.07 to 5.23 ± 0.01.

By contrast, the coupling-side objective changes only modestly: relative cost reduction increases from 3.91% at $M = 2$ to 3.96% at $M = 5$ and 3.97% at $M = 10$. Hence endpoint quadratic cost

Table 10: Sample standard deviation of CIFAR-10 FID with the Rectified Flow pretrained backbone, corresponding to Table 2. Standard deviations are computed over three seeds. For the RF backbone row, the standard deviation reflects sampling-seed variability; for the fine-tuned methods, it reflects variability across training seeds.

Method	NFE 1	NFE 2	NFE 4	NFE 8	NFE 10	NFE 50
RF (backbone)	0.166	0.291	0.174	0.115	0.094	0.011
Teacher Reflow (MSE)	0.010	0.021	0.026	0.022	0.023	0.021
One-step distillation (MSE)	0.006	0.246	0.359	0.362	0.360	0.393
Fresh-independent FT	2.154	1.828	0.247	0.136	0.091	0.040
OT-CFM-FT (Tong et al. (2023))	0.619	0.370	0.053	0.064	0.038	0.030
AlignFlow (Kong et al. (2026))	0.832	0.194	0.101	0.066	0.061	0.041
SD-FM (Mousavi-Hosseini et al. (2026))	0.202	0.251	0.087	0.034	0.029	0.034
Noise-ReCouple (cf. Hui et al. (2025))	0.226	0.099	0.121	0.135	0.121	0.147
ReCouple ($B = N$)	0.634	0.288	0.150	0.049	0.033	0.051
ReCouple	0.283	0.188	0.143	0.175	0.173	0.126
ReCouple-AS	0.608	0.202	0.121	0.126	0.153	0.067
ReCouple+	0.106	0.291	0.160	0.148	0.136	0.071
ReCouple++	0.748	0.008	0.081	0.124	0.113	0.081

Table 11: Sample standard deviation of CIFAR-10 FID with the OT-CFM pretrained backbone, corresponding to Table 3. Standard deviations are computed over three seeds. For the OT-CFM backbone row, the standard deviation reflects sampling-seed variability; for the fine-tuned methods, it reflects variability across training seeds.

Method	NFE 1	NFE 2	NFE 4	NFE 8	NFE 10	NFE 50
OT-CFM (backbone)	0.141	0.035	0.161	0.040	0.023	0.010
Teacher Reflow (MSE)	0.149	0.102	0.101	0.114	0.115	0.113
One-step distillation (MSE)	0.012	0.133	0.130	0.160	0.158	0.233
Fresh-independent FT	1.864	1.150	0.309	0.124	0.090	0.010
OT-CFM-FT (Tong et al. (2023))	0.596	0.287	0.142	0.110	0.121	0.059
AlignFlow (Kong et al. (2026))	1.078	0.462	0.172	0.075	0.054	0.052
SD-FM (Mousavi-Hosseini et al. (2026))	0.587	0.356	0.099	0.045	0.045	0.009
Noise-ReCouple (cf. Hui et al. (2025))	0.719	0.066	0.061	0.153	0.154	0.142
ReCouple ($B = N$)	0.278	0.216	0.151	0.101	0.075	0.005
ReCouple	0.156	0.061	0.104	0.129	0.136	0.084
ReCouple-AS	0.574	0.152	0.170	0.167	0.164	0.150
ReCouple+	0.669	0.252	0.095	0.060	0.030	0.016
ReCouple++	0.612	0.340	0.105	0.038	0.048	0.046

alone does not explain the large change in the FID profile. The results are instead consistent with the role of M as a source-resolution parameter: finer source populationization can improve generation at larger NFEs while being detrimental in the few-step regime.

Spatial regularization. Here, η specifies the initial regularization strength relative to the CFM loss: we calibrate λ_0 on a fixed batch so that $\lambda_0 \mathcal{L}_{\mathrm{AS}} / \mathcal{L}_{\mathrm{CFM}} = \eta$ at initialization. Thus, η is a loss ratio, rather than the regularization coefficient itself; $\eta = 0$ disables regularization.

The η sweep cleanly separates neural realization from coupling construction. All ReCouple++ settings use exactly the same $M = 5$ pooled plan, so the coupling diagnostics are unchanged as η varies. Nevertheless, increasing η monotonically improves mean FID from NFE 4 onward. Relative to unregularized ReCouple+ at $M = 5$, $\eta = 1$ reduces FID from 23.25 ± 0.16 to 22.75 ± 0.15 at NFE 4, from 12.61 ± 0.11 to 11.94 ± 0.17 at NFE 8, from 11.04 ± 0.11 to 10.33 ± 0.15 at NFE 10, and from 5.76 ± 0.04 to 5.00 ± 0.13 at NFE 50. Its effect at NFE 1–2 is comparatively small. Since the endpoint plan is fixed, these gains arise downstream of coupling construction, through how the transport field is learned and represented.

Table 12: CIFAR-10 ablations with the RF backbone at low NFEs. For ReCouple, we jointly vary the block size B and number of projection proposals R. For ReCouple+ and ReCouple++, we vary the pooling factor M and spatial-regularization strength η, respectively, while fixing the remaining hyperparameters to their default values. Each cell reports FID as mean $\pm$ sample standard deviation over three seeds (lower is better).

Method	Ablation	B	M	R	η	NFE 1	NFE 2	NFE 4
ReCouple	(B, R)	512	1	4	0	137.42 ± 1.04	48.89 ± 0.36	23.48 ± 0.13
		512	1	16	0	137.42 ± 1.04	48.89 ± 0.36	23.48 ± 0.13
		512	1	32	0	137.42 ± 1.04	48.89 ± 0.36	23.48 ± 0.13
		1024	1	4	0	130.07 ± 0.26	47.02 ± 0.05	22.99 ± 0.12
		1024	1	16	0	130.07 ± 0.26	47.02 ± 0.05	22.99 ± 0.12
		1024	1	32	0	130.07 ± 0.26	47.02 ± 0.05	22.99 ± 0.12
		2048	1	4	0	123.16 ± 0.28	45.72 ± 0.19	23.01 ± 0.14
		2048	1	16	0	123.16 ± 0.28	45.72 ± 0.19	23.01 ± 0.14
		2048	1	32	0	123.16 ± 0.28	45.72 ± 0.19	23.01 ± 0.14
ReCouple+	M	1024	2	16	0	155.74 ± 0.32	52.72 ± 0.05	21.38 ± 0.03
		1024	5	16	0	188.44 ± 0.39	66.81 ± 0.42	23.25 ± 0.16
		1024	10	16	0	195.09 ± 0.73	71.41 ± 0.07	24.59 ± 0.22
ReCouple++	η	1024	5	16	0.25	188.16 ± 0.58	66.59 ± 0.53	23.05 ± 0.12
		1024	5	16	0.50	187.91 ± 0.55	66.68 ± 0.17	22.95 ± 0.16
		1024	5	16	1.00	187.94 ± 0.55	66.56 ± 0.55	22.75 ± 0.15

Table 13: CIFAR-10 ablations with the RF backbone at higher NFEs. Experimental settings are identical to Table 12. Each cell reports FID as mean $\pm$ sample standard deviation over three seeds (lower is better).

Method	Ablation	B	M	R	η	NFE 8	NFE 10	NFE 50
ReCouple	(B, R)	512	1	4	0	16.99 ± 0.17	15.94 ± 0.20	11.49 ± 0.16
		512	1	16	0	16.99 ± 0.17	15.94 ± 0.20	11.49 ± 0.16
		512	1	32	0	16.99 ± 0.17	15.94 ± 0.20	11.49 ± 0.16
		1024	1	4	0	16.53 ± 0.08	15.45 ± 0.08	10.95 ± 0.11
		1024	1	16	0	16.53 ± 0.08	15.45 ± 0.08	10.95 ± 0.11
		1024	1	32	0	16.53 ± 0.08	15.45 ± 0.08	10.95 ± 0.11
		2048	1	4	0	16.62 ± 0.18	15.54 ± 0.17	10.95 ± 0.13
		2048	1	16	0	16.62 ± 0.18	15.54 ± 0.17	10.95 ± 0.13
		2048	1	32	0	16.62 ± 0.18	15.54 ± 0.17	10.95 ± 0.13
ReCouple+	M	1024	2	16	0	13.84 ± 0.01	12.68 ± 0.03	8.28 ± 0.07
		1024	5	16	0	12.61 ± 0.11	11.04 ± 0.11	5.76 ± 0.04
		1024	10	16	0	12.67 ± 0.02	10.93 ± 0.01	5.23 ± 0.01
ReCouple++	η	1024	5	16	0.25	12.40 ± 0.09	10.85 ± 0.11	5.49 ± 0.07
		1024	5	16	0.50	12.28 ± 0.09	10.68 ± 0.10	5.31 ± 0.08
		1024	5	16	1.00	11.94 ± 0.17	10.33 ± 0.15	5.00 ± 0.13

Takeaway. Taken together, the ablations separate four roles as follows. Increasing B reduces the endpoint-cost gap to empirical OT and is most beneficial in the few-step regime; increasing M instead induces a clear coarse-to-fine trade-off across NFEs; and spatial regularization can improve a fixed pooled coupling without altering its endpoint cost. Overall, the results show that endpoint transport cost alone is insufficient to characterize a useful coupling for flow matching: downstream quality depends on how coupling resolution, neural realization, and numerical resolution interact.

E.4 ImageNet-256 Implementation Details

All experiments use the class-conditional SiT-XL/2 model initialized from the same pretrained checkpoint and operate on ImageNet-256 MSE-VAE latents. Post-training uses AdamW with learn-

Table 14: Coupling diagnostics for the CIFAR-10 RF ablations. Relative cost reduction is $100(1 - C/C_{\mathrm{ind}})$, where C is the average sum of squared pixelwise endpoint differences and C_{ind} is the cost of independent pairing using the same source pool and empirical target marginal for each M. For $M = 1$, we additionally report the normalized endpoint-cost gap to global empirical OT, $(C - C_{\mathrm{OT}})/(C_{\mathrm{ind}} - C_{\mathrm{OT}})$. The corresponding global pooled OT optima were not computed for $M > 1$. For each B, $R \in \{4, 16, 32\}$ yields the same assignment; all ReCouple++ settings share the $M = 5$ ReCouple+ plan.

Method	Setting	Rel. cost reduction (%) ↑	Norm. empirical-OT gap ↓
ReCouple	$B = 512$	3.62	0.256
ReCouple	$B = 1024$	3.84	0.211
ReCouple	$B = 2048$	4.04	0.169
ReCouple+	$M = 2$	3.91	–
ReCouple+	$M = 5$	3.96	–
ReCouple+	$M = 10$	3.97	–

ing rate 10^{-4}, four warmup steps, cosine decay with minimum learning-rate ratio 0.1, gradient clipping at 1.0, and EMA decay 0.999. We use a physical batch size of 8 and eight gradient-accumulation steps, yielding an effective batch size of 64. Each coupling plan contains $6{,}000$ updates. Based on NFE-5 validation FID, we select the 2K ReCouple checkpoint, 3K ReCouple-AS and ReCouple++ checkpoints, 6K minibatch-OT checkpoint, and 3K SD-FM checkpoint.

ReCouple uses class-conditional coupling pools of size $4{,}096$. ReCouple++ uses pooling factor $M = 2$. Minibatch OT computes assignments within each batch of 64. For SD-FM, we solve one unregularized semi-discrete transport problem per ImageNet class using assignments

$$j^\star(z) = \arg\max_j \{\langle z, y_j \rangle + g_j\}.$$

The potentials are optimized with AdaGrad for $1{,}500$ steps using source batches of 256, with a constant schedule for the first $1{,}000$ steps and averaging over the final 250 steps. Assignment balance is estimated using $65{,}536$ source samples, after which we construct a fixed $6{,}000 \times 64$ pairing plan using assignment batches of 128.

For evaluation, we generate $50{,}000$ samples using Euler integration, classifier-free guidance scale 4.0, per-process batch size 8, four GPUs, and global seed 123. An evaluation at NFE K uses $K + 1$ time-grid points and K Euler updates. All metrics are computed against the same ImageNet-256 reference statistics using the guided-diffusion evaluator.

F Additional Empirical Results on ImageNet 256x256

Improved MeanFlow. We additionally evaluate ReCouple on an Improved MeanFlow model (iMF) (Geng et al., 2026) with a SiT-B/4 backbone on ImageNet 256×256. Starting from the same pretrained MeanFlow checkpoint, we apply ReCouple for 500 post-training steps and evaluate both models using identical sampling settings. We generate 50,000 samples with uniform Euler discretization and report FID, precision, and recall across $\mathrm{NFE} \in \{1, 2, 4, 5\}$.

As shown in Table 15, ReCouple consistently improves FID at every evaluated sampling budget. The improvement is already visible at one function evaluation and persists at larger NFE, reducing FID from 8.020 to 7.863 at NFE 5. ReCouple also improves precision across all budgets for which results are available. Recall also improves slightly at NFE 1–4, for which both methods are reported. These results indicate that the gains from coupling refinement extend beyond conventional flow matching models to MeanFlow parameterizations.

Table 15: ImageNet 256×256 results for Improved MeanFlow (iMF) with a SiT-B/4 backbone. Metrics are computed using 50,000 generated samples. Bold values denote the best result at each NFE.

Metric	Method	NFE 1	NFE 2	NFE 4	NFE 5
FID ↓	iMF	13.301	9.010	8.177	8.020
	iMF + ReCouple	**13.142**	**8.901**	**7.998**	**7.863**
Precision ↑	iMF	0.6723	0.7506	0.7673	0.7714
	iMF + ReCouple	**0.6772**	**0.7570**	**0.7716**	**0.7755**
Recall ↑	iMF	0.4575	0.4269	0.4209	0.4099
	iMF + ReCouple	**0.4591**	**0.4322**	**0.4245**	—

Recall Table. Table 16 shows that the ReCouple variants improve recall over the SiT baseline across the evaluated NFEs, indicating that the FID improvements are accompanied by improved sample coverage.

Table 16: Recall on ImageNet 256×256 using 50,000 generated samples. Higher is better.

Method	NFE 1	NFE 2	NFE 4	NFE 5	NFE 10	NFE 25	NFE 50
SiT baseline	0.00000	0.23710	0.23060	0.23840	0.27180	0.30320	0.32430
ReCouple (2k)	0.00030	0.34730	0.30080	0.30720	0.31210	0.34680	0.36500
ReCouple-AS (3k)	0.00050	**0.37410**	**0.36040**	**0.37660**	0.41230	**0.44560**	0.45360
ReCouple++ AS, M=2 (3k)	**0.00100**	0.34790	0.35170	0.36860	**0.41350**	0.44140	**0.45860**
Minibatch OT (6k)	0.00050	0.34220	0.33000	0.33080	0.36060	0.39030	0.39810