Multi-Agent Coordination via Support-Preserving Distillation

Sangmin Lee **Youngju Na** **Chanmi Lee** **Sung-eui Yoon**[†]
School of Computing, KAIST
{alex6095,yjna2907,chanmi99,sungeui}@kaist.ac.kr
[†]Corresponding author

https://alex6095.github.io/mosdot/

Abstract

Offline MARL increasingly relies on generative policies to model multimodal joint behavior, typically by distilling a centralized teacher into decentralized one-step actors under the CTDE. We identify a failure mode at the teacher training stage: standard flow-based teachers pair noise with replay targets independently, so nearby noise samples can be routed toward conflicting coordination modes. The teacher then produces samples between valid modes, and because the distillation loss regresses each local actor onto the conditional mean of the teacher's output given local input, this error is not absorbed but propagated to the student. To remove this teacher-side artifact, we propose *Mode-Support Semi-Discrete Optimal Transport* (MoSDOT), which summarizes multimodal replay into a finite mode support with prescribed capacities and uses conditional semi-discrete optimal transport to assign each noise sample to a single mode before teacher training. We additionally study a shared-randomness variant that uses a shared noise component at execution to expose the residual gap intrinsic to strict-product execution. On controlled diagnostics and offline MARL benchmarks, MoSDOT improves endpoint quality and routing consistency, particularly on datasets exhibiting multimodal joint behavior.

1 Introduction

Offline multi-agent reinforcement learning (MARL) aims to learn coordinated policies from a fixed dataset, without further environment interaction. The core difficulty is that coordination is defined over the *joint action* distribution of all agents: individually plausible per-agent actions may combine into unseen joint actions that violate safety or yield low value. This is especially problematic when the data contains multiple valid coordination modes, defined as distinct joint behavior patterns, where a unimodal policy averages alternatives into an invalid in-between action. Expressive generative policies based on diffusion and flow matching address this multimodality by modeling joint action as a multi-mode distribution [1, 2, 3], but their multi-step sampling can be costly at decision time and complicate value-guided actor extraction through critic gradients [4]. A recent line of work therefore trains a centralized generative *teacher* over joint actions and *distills* it into decentralized one-step *students*, which act locally and admit more direct value-guided extraction [2].

Distillation, however, exposes a gap between joint modeling and decentralized execution. The teacher conditions on the full joint context and a shared noise sample, whereas each local actor sees only its own context and independent noise. If the teacher's choice between coordination modes depends on information hidden from a local actor, the ℓ_2 distillation loss regresses to the conditional mean and collapses distinct modes into an invalid in-between action. Mode-preserving distillation therefore requires not only that the teacher cover the target modes, but also that those modes be *compatible with the information available to local actors*. This makes coherent teacher training critical: standard

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

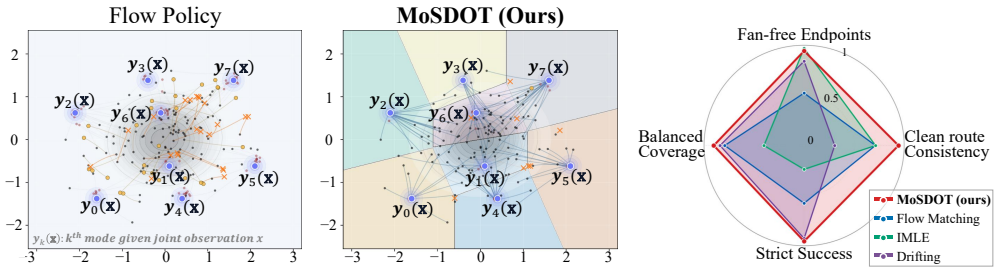

Figure 1: **Overview of source-to-mode routing. (Left) Standard Flow Policy:** independent noise–target pairing can entangle nearby source samples with incompatible coordination modes, producing off-support fan mass. **(Middle) MoSDOT:** semi-discrete optimal transport partitions noise space into capacity-matched cells assigned to coordinated behavior modes. **(Right)** Route quality is measured by off-support avoidance (*fan-free endpoints*), consistency of source-to-mode routing (*clean route consistency*), endpoint accuracy to valid modes (*strict success*), and balanced mass across modes (*balanced coverage*). Full metric definitions are provided in Appendix E.6.

flow-based training pairs noise samples with replay targets independently, leaving an unstructured mapping between noise and joint behaviors. With multimodal targets, adjacent noise points are forced to learn conflicting modes simultaneously, producing trajectories that pass through low-density regions or form spurious bridges between modes (Fig. 1, left). Because independent local policies capture only the conditional mean, even minor mode-mixing in the teacher can lead the student to collapse onto invalid joint actions. A clean teacher alone, however, is not sufficient. Independent agents cannot in general represent correlated joint support such as the anti-correlated XOR target, a residual mismatch we call the *product-projection gap*.

Our *Mode-Support Semi-Discrete Optimal Transport* (MoSDOT) addresses the teacher-side problem by summarizing multimodal replay into a finite set of representative joint behavior modes — the *mode support* — with prescribed capacities, and partitioning the noise space via semi-discrete optimal transport into cells assigned to those modes before teacher training (Fig. 1, middle). MoSDOT thus reduces teacher-side routing artifacts and preserves the mode support that local actors can recover, while leaving the product-projection gap explicit. To investigate this residual gap, we additionally consider a shared-randomness variant in which a portion of the noise input acts as an execution-time shared signal: all agents read the same coordination mode from this signal but still compute actions from their own observations and local noise, allowing correlated joint behavior without exchanging observations or actions.

In controlled diagnostics, MoSDOT mitigates teacher-side routing artifacts and improves source-to-mode assignment consistency. Fig. 1 (right) summarizes these gains across endpoint quality, route consistency, strict success, and coverage balance. On offline MARL benchmarks, MoSDOT performs competitively with existing generative baselines, particularly on datasets containing diverse joint behaviors. Overall, MoSDOT separates teacher-side source-routing artifacts from the residual product-projection gap, and provides a measurable interface between the centralized training and decentralized execution (CTDE) [5].

2 Related Work

2.1 Generative Policies and Distillation in Offline MARL

Offline reinforcement learning (RL) aims to learn policies from fixed datasets by constraining policy improvement close to actions represented in the data to avoid value overestimation [6, 7, 8, 9]. In multi-agent settings, the support is defined over coordinated *joint* actions, so small per-agent deviations can combine into joint behaviors absent from replay. Offline MARL methods address this through implicit constraints [10], actor rectification [11], and global-to-local value regularization under the CTDE. Standard actor or value-factorized parameterizations rarely model multimodal joint action support explicitly, so distinct coordination modes can be averaged into incoherent in-between actions.

Expressive generative policies address this multimodality. Diffusion and flow-based models have been used for trajectory-level planning [12, 13], action generation [14], and one-step offline policy learn-

ing [4]. In offline MARL, MADiff [15] models coordinated multi-agent trajectories with diffusion, and DoF [1] factorizes diffusion generation and formulates Individual-Global-identically-Distributed (IGD) generation as a distributional analogue of value factorization. These approaches improve expressiveness and execution efficiency, but the quality of the centralized generative distribution remains an important prerequisite.

Policy distillation transforms multimodal teachers into efficient one-step students. Since pointwise regression tends to average modes, recent work has introduced set-level objectives based on Implicit Maximum Likelihood Estimation (IMLE) and Chamfer matching [16], as well as drifting objectives that shape support through attraction toward data and repulsion among generated samples [17]. In offline RL, FQL [4] distills a behavioral flow into a Q-regularized one-step policy, and MAC-Flow [2] distills a centralized multi-agent flow teacher into decentralized one-step actors. These methods primarily focus on the distillation process itself, while our work instead studies the upstream construction of the teacher distribution before distillation begins.

2.2 Mode Alignment in Flow Policies and Factorized Execution

Flow-based generative models are often sensitive to how source noise is coupled to data during training. Flow matching trains continuous normalizing flows by regressing vector fields along conditional probability paths [18]. With independent noise–target pairings, multimodal targets can produce samples that fall between valid modes. Optimal transport (OT) flow matching [19] mitigates this effect by coupling noise and data through transport plans, and AlignFlow [20] uses semi-discrete optimal transport (SDOT) to partition source space into Laguerre cells assigned to empirical data samples for image generation.

A related issue arises in MARL, where centralized training must be compatible with decentralized execution. Value factorization methods such as VDN [21], QMIX [22], and QTRAN [23] impose structure on centralized value functions so that decentralized action selection remains aligned with a global objective. Generative offline MARL raises a distributional analogue of this question. A centralized teacher may represent multimodal joint behavior, but the distilled local actors see only local information and are often trained with pointwise regression losses.

Our work focuses on the corresponding upstream failure mode. Before student training, independent teacher pairings can create off-support or entangled source routes between coordinated joint modes. During distillation, mode variation that is not locally observable is further averaged by the student objective. MoSDOT reduces this teacher-side routing artifact by aligning source-noise regions with the mode support via conditional SDOT, providing cleaner teacher targets for support-preserving decentralized distillation.

3 Method

3.1 Problem Setup

We consider cooperative offline MARL under a decentralized partially observable Markov decision process (Dec-POMDP) [24], defined as $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{O}, \mathcal{A}, P, Z, R, \gamma \rangle$. Let $\mathcal{N} = \{1, ..., N\}$ be the set of N agents, with $\mathcal{S}$, $\mathcal{O} = \{\mathcal{O}_i\}_{i=1}^N$, and $\mathcal{A} = \{\mathcal{A}_i\}_{i=1}^N$ denoting the joint global state, observation, and action spaces, respectively. We denote $\mathbf{x} = (o_1, \ldots, o_N) \in \mathcal{O}$ as the joint observation context and $\mathbf{a} = (a_1, \ldots, a_N) \in \mathcal{A}$ as the joint action. $P(s'|s, \mathbf{a}) : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition function and $Z^i(o_i|s) : \mathcal{S} \to \Delta(\mathcal{O}_i)$ is the i-th agent's observation function, where $s \in \mathcal{S}$ and $\Delta(\cdot)$ denotes the set of probability measures over the given space. The team reward is generated based on the reward function $R(s, \mathbf{a}, s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and γ is a discount factor.

At each timestep, all decentralized agents choose their actions $\mathbf{a}$ conditioned on their local observations $\mathbf{x}$ provided by the environment in state s. The environment then transitions to $s' \sim P(s'|s, \mathbf{a})$, and the team reward is given by $R(s, \mathbf{a}, s')$. In the offline setting, learning uses a fixed dataset $\mathcal{D}$ of joint transitions collected by an unknown behavior policy, without further environment interaction.

Our goal is to learn each agent's policy as a decentralized actor through centralized training. During training, we first use a centralized generative teacher that has access to the joint context $\mathbf{x}$. The teacher is represented by a deterministic sampler conditioned on source noise distribution:

$$\tilde{\mathbf{a}} = \mu_\psi(\mathbf{x}, \mathbf{z}), \qquad \mathbf{z} = (z_1, \ldots, z_N) \sim \rho_0, \tag{1}$$

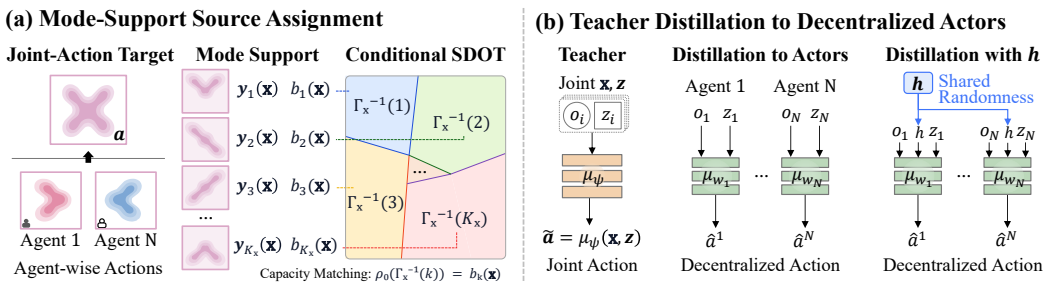

Figure 2: **MoSDOT-based framework for multi-agent distillation.** **(a)** Replay joint actions are summarized into a mode support $\{\mathbf{y}_k(\mathbf{x})\}$ with capacities $\{b_k(\mathbf{x})\}$, and conditional SDOT partitions the noise space into Laguerre cells $\Gamma_{\mathbf{x}}^{-1}(k)$ assigned to each mode. **(b)** The centralized teacher μ_ψ is distilled into decentralized one-step actors $\{\mu_{w_i}\}$, each using only its local observation and noise. An optional shared signal h correlates local noise to mitigate the strict-product gap.

where ρ_0 is a simple source noise distribution, such as a standard Gaussian. Since $\mathbf{z}$ is random, this sampler induces a conditional distribution over joint actions given $\mathbf{x}$. In the teacher stage, both the full joint observation context $\mathbf{x}$ and the full joint noise $\mathbf{z}$ may be used. At execution time, however, each agent must act using only its local information. We therefore deploy a strict-product policy:

$$\pi_{\mathbf{w}}^{\text{prod}}(\mathbf{a} \mid \mathbf{x}) = \prod_{i=1}^{N} \pi_{w_i}(a_i \mid o_i), \tag{2}$$

where $\mathbf{w} = (w_1, \ldots, w_N)$ denotes the set of decentralized actor parameters across N agents. In the one-step reparameterized implementation used throughout this work, the deployed action is $\hat{\mathbf{a}} = (\mu_{w_1}(o_1, z_1), \ldots, \mu_{w_N}(o_N, z_N))$, with each actor receiving only its local observation o_i and local noise block z_i, unless specified otherwise. Thus, the centralized teacher can model coordinated joint behavior using $\mathbf{x}$ and $\mathbf{z}$, while the student must reproduce such behavior under this decentralized information constraint.

For value-guided actor extraction, we use a centralized critic $Q_{\text{tot}}(\mathbf{x}, \mathbf{a})$ that satisfies the Individual-Global-Max (IGM) principle [21, 22, 23]:

$$\arg\max_{\mathbf{a}} Q_{\text{tot}}(\mathbf{x}, \mathbf{a}) \;=\; \Big(\arg\max_{a_1} Q_1(o_1, a_1),\; \ldots,\; \arg\max_{a_N} Q_N(o_N, a_N)\Big). \tag{3}$$

Here, $Q_{\text{tot}}(\mathbf{x}, \mathbf{a})$ denotes the global Q-function over the joint context and joint action, while $Q_i(o_i, a_i)$ denotes the individual utility for agent i. The IGM principle ensures that greedy selection over each local Q_i remains consistent with the global optimum of Q_{tot}.

The remainder of Sec. 3 describes how we train the teacher so that its source noise is assigned to coordinated joint-action modes, and how we distill the resulting teacher into decentralized actors.

3.2 Mode Support and Capacities

Learning the teacher sampler $\mu_\psi(\mathbf{x}, \mathbf{z})$ requires pairing source noise $\mathbf{z}$ with target joint actions so that the induced conditional distribution matches the target behavior at $\mathbf{x}$. A naive approach pairs $\mathbf{z}$ and the replay target without considering their relationship, mapping them arbitrarily. When the targets at a given $\mathbf{x}$ contain multiple distinct coordinated joint behaviors, such as two valid joint actions in a coordination game, random pairing entangles these behaviors in source-noise space. This often causes nearby noise vectors to map to unrelated joint actions, and the noise mass allocated to each behavior is determined by random sampling rather than designed intentionally.

We instead organize the targets at each $\mathbf{x}$ into a finite *mode support*, a discrete set of representative joint actions, and assign a *capacity* (a target source-noise mass) to each element of this support before training. The sampler then learns a structured assignment in which contiguous regions of $\mathbf{z}$-space correspond to single modes, with region sizes set by the capacities.

Mode support. For each conditioning context $\mathbf{x}$, we collect a finite set of $M_{\mathbf{x}}$ target joint actions:

$$\mathcal{T}_{\mathbf{x}} = \{\mathbf{a}^{(m)}(\mathbf{x})\}_{m=1}^{M_{\mathbf{x}}}, \qquad \mathbf{a}^{(m)}(\mathbf{x}) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_N. \tag{4}$$

In the default construction, these targets are replay actions from datasets. A support aggregation rule

$$q_{\mathbf{x}} : \{1, \dots, M_{\mathbf{x}}\} \to \{1, \dots, K_{\mathbf{x}}\} \tag{5}$$

groups the targets into $K_{\mathbf{x}}$ clusters. Each cluster k collects the target indices $C_k(\mathbf{x}) = \{m : q_{\mathbf{x}}(m) = k\}$ and is summarized by a representative joint action $\mathbf{y}_k(\mathbf{x}) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$, which we call a *mode*. The *mode support* at context $\mathbf{x}$ is the discrete set of these representatives:

$$\mathcal{Y}_{\mathbf{x}} = \{\mathbf{y}_1(\mathbf{x}), \dots, \mathbf{y}_{K_{\mathbf{x}}}(\mathbf{x})\}. \tag{6}$$

The support size $K_{\mathbf{x}}$ is determined by the observed targets and the aggregation rule, not by enumerating the full combinatorial joint action space. The aggregation rule $q_{\mathbf{x}}$ determines the granularity of $\mathcal{Y}_{\mathbf{x}}$. We use identity support for discrete-action benchmarks and a simple joint-action quantization for continuous-action benchmarks.

Capacities. Each mode $\mathbf{y}_k(\mathbf{x})$ is assigned a *capacity* $b_k(\mathbf{x}) \geq 0$ satisfying $\sum_{k=1}^{K_{\mathbf{x}}} b_k(\mathbf{x}) = 1$, specifying its target source-noise mass. We use $b_k(\mathbf{x}) = |C_k(\mathbf{x})|/M_{\mathbf{x}}$ so that the target source-noise mass matches the empirical replay frequency of each mode. Although each mode decomposes as $\mathbf{y}_k(\mathbf{x}) = \big(y_{k,1}(\mathbf{x}), \dots, y_{k,N}(\mathbf{x})\big)$, source assignment is performed on $\mathbf{y}_k(\mathbf{x})$ as a whole, so source-noise regions are assigned to coordinated joint behaviors rather than to a product of per-agent action modes. This is what allows the sampler to represent correlated joint behaviors that a per-agent factorization cannot.

3.3 Source Assignment via Semi-Discrete Optimal Transport

Sec. 3.2 defines the target mass $b_k(\mathbf{x})$ for each mode, but leaves open which source-noise samples should be paired with which mode. We resolve this by partitioning source-noise space so that each cell carries the prescribed mass and is assigned to a single mode. We adopt the SDOT-based source–target alignment introduced for image generation by Kong et al. [20], and apply it to the capacity-weighted mode support $\mathcal{Y}_{\mathbf{x}}$.

Let ρ_0 denote the source law over joint noise $\mathbf{z} = (z_1, \dots, z_N)$. We define the discrete target measure at context $\mathbf{x}$ as follows:

$$\nu_{\mathbf{x}} = \sum_{k=1}^{K_{\mathbf{x}}} b_k(\mathbf{x})\, \delta_{\mathbf{y}_k(\mathbf{x})}, \tag{7}$$

where $\delta_{\mathbf{y}_k(\mathbf{x})}$ denotes a Dirac measure at mode $\mathbf{y}_k(\mathbf{x})$, so $\nu_{\mathbf{x}}$ places mass $b_k(\mathbf{x})$ on each mode in $\mathcal{Y}_{\mathbf{x}}$. For a cost $c_{\mathbf{x}}(\mathbf{z}, k)$ between source noise $\mathbf{z}$ and mode $\mathbf{y}_k(\mathbf{x})$, the SDOT primal is defined as follows:

$$\min_{\Gamma_{\mathbf{x}}} \mathbb{E}_{\mathbf{z}\sim\rho_0}\big[c_{\mathbf{x}}\big(\mathbf{z}, \Gamma_{\mathbf{x}}(\mathbf{z})\big)\big] \quad \text{s.t.} \quad \rho_0\big(\Gamma_{\mathbf{x}}^{-1}(k)\big) = b_k(\mathbf{x}), \quad k = 1, \dots, K_{\mathbf{x}}. \tag{8}$$

Here, $\Gamma_{\mathbf{x}} : \mathbf{z} \mapsto k$ is the source-to-mode assignment, mapping each source sample to a mode index. We take $c_{\mathbf{x}}$ to be the squared Euclidean distance in the representation that defines $\mathcal{Y}_{\mathbf{x}}$. The SDOT dual is a finite-dimensional concave problem and admits stochastic optimization on samples from ρ_0 [25]. We write its objective as:

$$\Phi_{\mathbf{x}}(g) = \sum_{k=1}^{K_{\mathbf{x}}} b_k(\mathbf{x}) g_k + \mathbb{E}_{\mathbf{z}\sim\rho_0}\left[\min_{j\in\{1,\dots,K_{\mathbf{x}}\}} \big(c_{\mathbf{x}}(\mathbf{z}, j) - g_j\big)\right]. \tag{9}$$

The optimal dual weights are obtained as $g_{\mathbf{x}}^\star \in \arg\max_{g\in\mathbb{R}^{K_{\mathbf{x}}}} \Phi_{\mathbf{x}}(g)$. The resulting Laguerre cells define the source-to-mode assignment:

$$\Gamma_{\mathbf{x}}(\mathbf{z}) = \arg\min_k \big[c_{\mathbf{x}}(\mathbf{z}, k) - g_{x,k}^\star\big]. \tag{10}$$

We solve Eq. 9 stochastically with Monte-Carlo sampling from ρ_0. For each pair $(\mathbf{x}, \mathbf{z})$, the assignment $\Gamma_{\mathbf{x}}$ selects the mode index, and we write the corresponding mode as $\mathbf{y}_\Gamma(\mathbf{x}, \mathbf{z}) = \mathbf{y}_{\Gamma_{\mathbf{x}}(\mathbf{z})}(\mathbf{x})$. We then form the linear interpolation path $\bar{\mathbf{a}}_t = (1-t)\mathbf{z} + t\,\mathbf{y}_\Gamma(\mathbf{x}, \mathbf{z})$ with displacement $\mathbf{v}_\Gamma(\mathbf{x}, \mathbf{z}) = \mathbf{y}_\Gamma(\mathbf{x}, \mathbf{z}) - \mathbf{z}$, where $t \sim \mathrm{Unif}([0, 1])$. Following [18], the teacher learns a velocity field u_ψ by regressing onto the assigned displacement:

$$\mathcal{L}_{\text{teacher}}(\psi) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\, \mathbf{z}\sim\rho_0,\, t\sim\mathrm{Unif}([0,1])}\left[\big\|u_\psi(\bar{\mathbf{a}}_t, t, \mathbf{x}) - \mathbf{v}_\Gamma(\mathbf{x}, \mathbf{z})\big\|_2^2\right]. \tag{11}$$

At inference, starting from $\mathbf{z}$ and integrating the learned field u_ψ from $t = 0$ to $t = 1$ defines the teacher sampler $\mu_\psi(\mathbf{x}, \mathbf{z})$ used in Sec. 3.1. Unlike independent pairing, this couples source regions to modes with capacity-controlled mass.

3.4 Joint Policy Distillation

The teacher trained in Sec. 3.3 produces coordinated joint-action samples using the full context and joint noise. We distill these rollouts into the strict-product student family in Eq. (2). For a replay context $\mathbf{x} = (o_1, \dots, o_N)$ and source noise $\mathbf{z} = (z_1, \dots, z_N) \sim \rho_0$, the teacher and student actions are defined as:

$$\tilde{\mathbf{a}} = \mu_\psi(\mathbf{x}, \mathbf{z}), \qquad \hat{\mathbf{a}} = \big(\mu_{w_1}(o_1, z_1), \dots, \mu_{w_N}(o_N, z_N)\big). \tag{12}$$

The distillation loss matches each local actor to the corresponding component of the centralized teacher rollout:

$$\mathcal{L}_{\text{distill}}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\, \mathbf{z}\sim\rho_0}\left[\sum_{i=1}^{N} \|\mu_{w_i}(o_i, z_i) - [\mu_\psi(\mathbf{x}, \mathbf{z})]_i\|_2^2\right], \tag{13}$$

where $[\cdot]_i$ extracts the action component for agent i. For value-guided actor extraction, we optimize:

$$\mathcal{L}_\pi(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\, \mathbf{z}\sim\rho_0}\left[-Q_{\text{tot}}(\mathbf{x}, \hat{\mathbf{a}}) + \alpha \sum_{i=1}^{N} \|\mu_{w_i}(o_i, z_i) - [\mu_\psi(\mathbf{x}, \mathbf{z})]_i\|_2^2\right], \tag{14}$$

where $\hat{\mathbf{a}}$ is as defined in Eq. (12), Q_{tot} is the IGM-compatible centralized critic of Eq. (3), and α controls the strength of teacher regularization. The assignment $\Gamma_{\mathbf{x}}$ is used only to train the teacher and is not given to the decentralized actors at execution time.

Strict-product gap. MoSDOT improves the centralized teacher, but it does not remove the execution-time factorization constraint. A strict-product actor cannot represent every correlated joint distribution. For example, even when MoSDOT places clean teacher mass on $(+1, -1)$ and $(-1, +1)$, a strict-product student with the same local marginals also assigns mass to unsupported tuples $(+1, +1)$ and $(-1, -1)$: each agent independently samples its own noise, so the two cannot agree on which mode to execute. We call this residual mismatch the *strict-product gap*.

The MoSDOT construction itself suggests a way to probe this gap. Recall that the assignment $\Gamma_{\mathbf{x}}$ partitions the noise space into Laguerre cells, each mapped to a single mode $\mathbf{y}_k(\mathbf{x})$. If all agents observe the same cell index, they automatically agree on the same mode and avoid unsupported product tuples. This motivates a shared-randomness variant. Let $h \sim \rho_h$ be a random signal shared across agents at execution, playing the role of a shared cell index, and let $\epsilon_i \sim \rho_i$ be agent-local noise for within-mode variability. The resulting policy can be modified as:

$$\pi^{\text{sr}}_{\mathbf{w}}(\mathbf{a} \mid \mathbf{x}) = \mathbb{E}_{h\sim\rho_h}\left[\prod_{i=1}^{N} \pi^{\text{sr}}_{w_i}(a_i \mid o_i, h)\right], \tag{15}$$

with one-step implementation $\hat{\mathbf{a}}^{\text{sr}} = \big(\mu^{\text{sr}}_{w_1}(o_1, h, \epsilon_1), \dots, \mu^{\text{sr}}_{w_N}(o_N, h, \epsilon_N)\big)$. Without h, this reduces to strict-product execution as in Eq. 2. With h, agents act locally and exchange no observations, actions, or messages, but their mode choice is correlated through the shared signal. This isolates the residual error due to independent local sampling and complements the teacher-side coupling provided by MoSDOT. Further theoretical and algorithmic details are provided in Appendices A.5 and B.2.

4 Experiments

Our experiments separate three sources of error in CTDE generative distillation: (i) teacher-side off-support mass, (ii) incompatibility between the teacher's source assignment and decentralized actors, and (iii) the residual strict-product gap. MoSDOT is designed to address first two by replacing independent source–target pairing with a capacity-matched SDOT assignment. We then compare strict-product and shared-randomness decentralized actors to isolate what can be preserved under decentralized execution.

4.1 Experimental Setup

Diagnostics and benchmarks. We use two controlled tasks to visualize the failure modes studied in this work. The landmark diagnostic in Fig. 3 evaluates teacher-side source assignment with multiple coordinated modes across three agents. The XOR task in Fig. 5, adapted from MAC-Flow [2], isolates the strict-product gap using the anti-correlated support $\{(-1, +1), (+1, -1)\}$. For benchmark evaluation, we use SMACv1 [26], SMACv2 [27], and MPE [28].

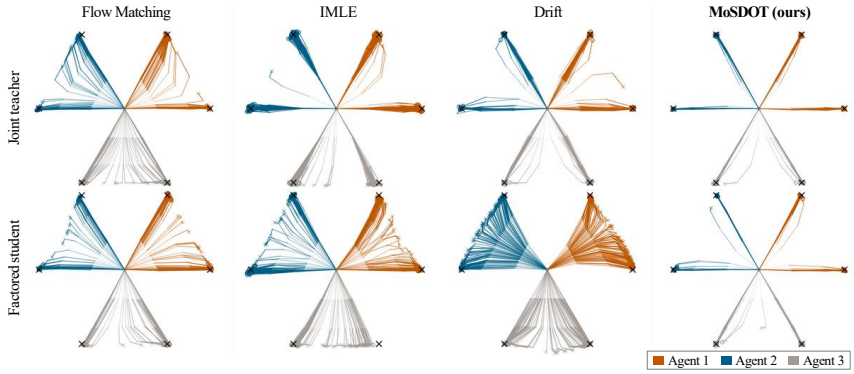

Figure 3: **Landmark policy trajectories.** Black $\times$ marks denote the six anchor mode endpoints, and colors indicate the three anchor pairs. The top row shows centralized joint teacher rollouts, and the bottom row shows one-step factorized students distilled under independent local noise. Baseline students show fan mass between modes, while MoSDOT reduces fan mass and retains the structure.

Method	Fan-free ↑	Route cons. ↑	Success ↑	Balance ↑
Flow matching	0.857 $\pm$0.010	0.856 $\pm$0.010	0.915 $\pm$0.029	0.917 $\pm$0.043
IMLE	**0.991** $\pm$0.007	0.857 $\pm$0.023	0.848 $\pm$0.064	0.761 $\pm$0.029
Drifting	0.952 $\pm$0.005	0.653 $\pm$0.023	0.985 $\pm$0.002	0.936 $\pm$0.029
MoSDOT	0.983 $\pm$0.002	**0.972** $\pm$0.001	**0.991** $\pm$0.002	**0.961** $\pm$0.011

Table 1: **Joint-teacher landmark diagnostics.** Mean $\pm$ one standard deviation over 4 seeds.

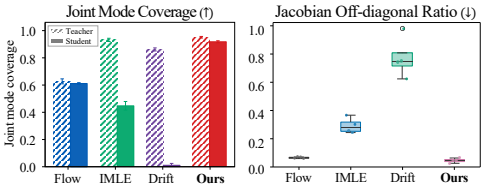

Figure 4: **Landmark route-locality diagnostics.**

Baselines and protocol. For controlled tasks, we compare against independent flow matching [18], IMLE policy [29], and Drifting model [17]. For benchmark evaluation, we follow the protocol and baseline suite of MAC-Flow [2], which includes representative Gaussian, diffusion, and flow-policy baselines such as MADiff, DoF, and MAC-Flow [15, 1, 2]. Unless stated otherwise, benchmark results use 6 seeds and report two standard deviations after $\pm$. More details on baselines and implementations are provided in Appendix E.

Diagnostic metrics. For the landmark diagnostic, we evaluate teacher quality along two axes. **(i) Endpoint support**: does the teacher place samples on the modes in $\mathcal{Y}_{\mathbf{x}}$? *Fan-free* counts actions inside accepted directional cones. *Success* measures landmark reach rate, and *Balance* measures balance between valid local modes. **(ii) Source routing**: how does the teacher use source noise to choose among modes? *Route consistency* combines fan-free mass with the stability of mode labels among nearby samples in an agent's noise slice. *Jacobian off-diagonal ratio* measures how strongly the teacher's output for agent i depends on other agents' noise $z_{j\neq i}$ relative to its own noise z_i, with values near zero indicating weak local cross-noise sensitivity. Endpoint metrics test whether the teacher reaches $\mathcal{Y}_{\mathbf{x}}$, and routing metrics test whether mode choices depend only on locally observable information. Full definitions are given in Appendix E.6.

4.2 Endpoint Accuracy Does Not Imply Distillable Assignments

We first test whether reaching the correct modes is sufficient for factorized distillation. Tab. 1 measures endpoint-level support quality, while Fig. 4 asks a complementary question about whether the teacher's mode choice for an agent depends on the agent's own noise or the noise of other agents.

A teacher can reach valid modes while choosing among them in a way unavailable to the corresponding local actor. Flow matching has reasonable route consistency and balance, but its lower fan-free score reveals samples between modes. IMLE achieves the best fan-free score with sharp endpoints, but sacrifices success and balance, appearing as uneven mass across modes in Fig. 4. Drifting shows this mismatch most strongly. It reaches valid modes often, but has the lowest route consistency in Tab. 1 and the largest Jacobian off-diagonal ratio in Fig. 4, indicating its per-agent output is strongly shaped by other agents' noise—information unavailable to the local actor. Endpoint support is therefore insufficient: when mode choice depends on off-agent noise, distillation can average or mix modes even with accurate endpoints. We evaluate teacher quality along both axes—whether it reaches $\mathcal{Y}_{\mathbf{x}}$, and whether mode choices are reproducible from locally available information.

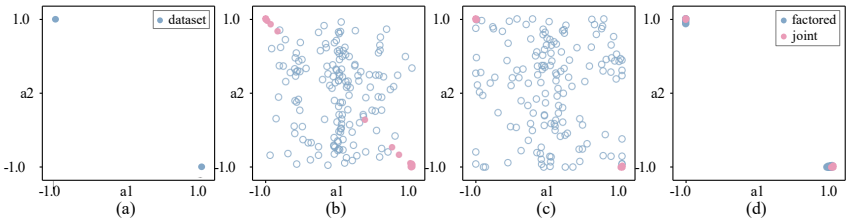

Figure 5: **XOR strict-product visualization. (a)** Target support placing equal mass on $(-1, +1)$ and $(+1, -1)$. **(b) MAC-Flow**: bridge mass spreads across all quadrants under strict-product execution. **(c) Ours w/o shared signal**: clean teacher, residual strict-product gap. **(d) Ours with shared signal**: gap closed. Blue: factored student; pink: joint teacher.

4.3 MoSDOT Preserves Support Without Entangling Assignments

The previous section showed that no baseline is strong on both axes simultaneously. We now evaluate whether MoSDOT closes this gap. On the same landmark replay, MoSDOT replaces independent source–target pairing with the capacity-matched SDOT assignment of Eq. (11), using replay frequencies as mode capacities.

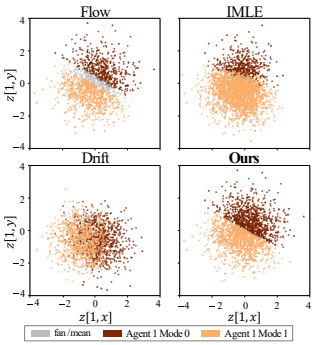

Figure 6: Source-space partitions in Agent 1's noise plane.

MoSDOT improves support sharpness without sacrificing assignment locality. On the endpoint axis (Tab. 1), it is competitive with the best endpoint-focused baselines on fan-free support and success, and achieves the best balance. The teacher's joint mode coverage in Fig. 4 (left, hatched) is also the highest across methods. On the routing axis, its route consistency is the highest among all methods, and its Jacobian off-diagonal ratio in Fig. 4 (right) is comparable to flow matching, well below IMLE and Drifting. Fig. 6 visualizes both axes jointly in Agent 1's noise plane: Flow leaves a fan between modes, IMLE assigns unbalanced area, and Drift mixes the two modes throughout, while MoSDOT yields balanced, fan-less, disentangled regions.

4.4 Cleaner Source Assignments Survive Decentralized Distillation

Teacher-side improvement is useful only if it persists after decentralized distillation. To examine the teacher-to-student transfer, we distill strict-product students from Flow Matching, IMLE, Drifting, and MoSDOT teachers under the same student architecture, seed list, aggregation rule, and training budget. MoSDOT uses the capacity-matched teacher from Sec. 4.3.

The student appears to track the teacher's geometry. Fig. 3 (bottom row) shows that the MoSDOT student largely follows the teacher's ray structure, while baselines show fan mass between modes. Quantitatively, the MoSDOT-distilled student retains higher fan-free, success, and balance scores than baselines, while IMLE and Drifting students lose substantial endpoint quality through distillation (Appendix F.1). These results suggest that improving the centralized source assignment can improve the support preserved by decentralized actors.

4.5 Shared Randomness Isolates the Strict-Product Gap

Even with a clean teacher, the strict-product student of Eq. (2) cannot in general represent correlated joint support. The XOR target $\{(-1, +1), (+1, -1)\}$ illustrates this: independent local sampling that matches the marginals also produces the unsupported tuples $(+1, +1)$ and $(-1, -1)$. We use the shared-randomness variant of Eq. (15) to separate this product-projection error from teacher-side assignment error. Fig. 5 shows the separation across three stages. The unrouted flow teacher places bridge mass between the two valid tuples, which a strict-product student spreads across all four quadrants (b). MoSDOT concentrates the teacher mass on the two valid tuples, yet independent local sampling still mixes the two modes after distillation (c). Shared randomness, with the same MoSDOT teacher, lets the actors agree on a mode-selection signal while still acting locally, and the student concentrates on the valid tuples (d). The transition from (c) to (d)—where only the execution-time signal changes—isolates the strict-product gap from teacher-side assignment failure.

Table 2: **Performance on discrete-action offline MARL benchmarks (SMACv1 and SMACv2).** Mean over 6 seeds, $\pm 2\sigma$. Best and second-best in sage and powder blue.

Benchmark	Scenario	Dataset	Offline / BC baselines			Diffusion policies			Flow / OT policies		
			BC	MABCQ	MACQL	Diffusion BC	MADiff	DoF	Flow BC	MAC-Flow	MoSDOT
SMACv1	3m	Good	16.0 ± 1.0	3.7 ± 1.1	19.1 ± 0.1	19.5 ± 0.5	19.3 ± 0.5	19.8 ± 0.2	$\mathbf{20.0 \pm 0.0}$	19.8 ± 0.2	$\mathbf{20.0 \pm 0.0}$
		Medium	8.2 ± 0.8	4.0 ± 1.0	13.7 ± 0.3	13.3 ± 0.7	16.4 ± 2.6	18.6 ± 1.2	14.7 ± 1.5	18.0 ± 3.2	$\mathbf{18.8 \pm 1.5}$
		Poor	4.4 ± 0.1	3.4 ± 1.0	4.2 ± 0.1	4.2 ± 0.2	10.3 ± 6.1	10.9 ± 1.1	4.5 ± 0.1	10.6 ± 2.2	$\mathbf{16.4 \pm 4.9}$
	8m	Good	16.7 ± 0.4	4.8 ± 0.6	18.9 ± 0.9	19.4 ± 0.5	18.9 ± 1.1	19.6 ± 0.3	19.5 ± 0.2	19.7 ± 0.3	$\mathbf{20.0 \pm 0.0}$
		Medium	10.7 ± 0.5	5.6 ± 0.6	15.5 ± 1.5	18.6 ± 0.6	16.8 ± 1.6	18.6 ± 0.8	18.2 ± 0.8	19.4 ± 0.6	$\mathbf{20.0 \pm 0.0}$
		Poor	5.3 ± 0.1	3.6 ± 0.8	7.5 ± 1.0	4.8 ± 0.2	9.8 ± 0.9	$\mathbf{12.0 \pm 1.2}$	4.9 ± 0.1	11.5 ± 0.8	10.8 ± 0.8
	2s3z	Good	18.2 ± 0.4	7.7 ± 0.9	17.4 ± 0.3	18.0 ± 1.0	15.9 ± 1.2	18.5 ± 0.8	19.5 ± 0.1	19.5 ± 0.5	$\mathbf{20.1 \pm 0.1}$
		Medium	12.3 ± 0.7	7.6 ± 0.7	15.6 ± 0.4	13.4 ± 1.4	15.6 ± 0.3	18.1 ± 0.9	15.1 ± 2.0	17.6 ± 0.6	$\mathbf{18.2 \pm 1.0}$
		Poor	6.7 ± 0.3	6.6 ± 0.2	8.4 ± 0.8	6.2 ± 1.2	8.5 ± 1.3	$\mathbf{10.0 \pm 1.1}$	6.9 ± 0.8	8.5 ± 0.6	9.4 ± 0.8
	5m_vs_6m	Good	15.8 ± 3.6	2.4 ± 0.4	16.2 ± 1.6	16.8 ± 2.3	16.5 ± 2.8	17.7 ± 1.1	14.7 ± 2.1	$\mathbf{18.6 \pm 3.5}$	18.5 ± 1.9
		Medium	12.4 ± 0.9	3.8 ± 0.5	15.1 ± 2.9	12.5 ± 2.1	15.2 ± 2.6	16.2 ± 0.9	12.8 ± 0.8	15.6 ± 1.3	$\mathbf{17.9 \pm 1.2}$
		Poor	7.5 ± 0.2	3.3 ± 0.5	10.5 ± 3.1	8.0 ± 1.0	8.9 ± 1.3	10.8 ± 0.3	7.7 ± 0.8	9.8 ± 2.1	$\mathbf{12.3 \pm 0.9}$
	2c_vs_64zg	Good	17.5 ± 0.4	10.1 ± 0.2	12.9 ± 0.2	17.8 ± 1.3	14.7 ± 2.2	16.1 ± 0.8	18.0 ± 1.3	19.1 ± 0.8	$\mathbf{20.2 \pm 0.6}$
		Medium	12.5 ± 0.3	9.9 ± 0.2	11.6 ± 0.1	10.5 ± 1.1	12.8 ± 1.2	13.9 ± 0.9	11.8 ± 2.6	14.9 ± 4.1	$\mathbf{16.4 \pm 1.0}$
		Poor	9.7 ± 0.2	9.0 ± 0.2	10.2 ± 0.1	10.2 ± 2.3	10.8 ± 1.1	$\mathbf{11.5 \pm 1.1}$	10.0 ± 0.3	11.4 ± 0.4	11.4 ± 1.4
	Average rewards		12.2	5.5	13.1	13.0	13.8	15.6	13.4	15.6	**16.7**
SMACv2	terran_5_vs_5	replay	7.3 ± 1.0	13.8 ± 4.4	11.8 ± 0.9	9.3 ± 0.9	13.3 ± 1.8	15.4 ± 1.3	8.3 ± 1.9	$\mathbf{16.6 \pm 4.3}$	$\mathbf{16.6 \pm 4.3}$
	zerg_5_vs_5	replay	6.8 ± 0.6	10.3 ± 1.2	10.3 ± 3.4	8.1 ± 1.7	10.2 ± 1.1	$\mathbf{12.0 \pm 1.1}$	4.6 ± 0.5	9.8 ± 1.5	9.7 ± 2.8
	terran_10_vs_10	replay	7.4 ± 0.5	12.7 ± 2.0	11.8 ± 2.0	5.5 ± 1.5	13.8 ± 1.3	$\mathbf{14.6 \pm 1.1}$	5.8 ± 1.7	13.0 ± 4.7	9.4 ± 2.6
	Average rewards		7.2	12.3	11.3	7.6	12.4	**14.0**	6.2	13.1	11.9

Table 3: **Performance on continuous-action offline MARL benchmarks (MPE Simple Spread).** *M-R* denotes Medium-Replay.

Scenario	Dataset	Offline MARL					Generative policies		
		MATD3BC	MACQL	ICQ	OMAR	OMIGA	MADiff	MAC-Flow	MoSDOT
Spread	Expert	108.3 ± 3.3	98.2 ± 5.2	$\mathbf{114.9 \pm 2.6}$	104.0 ± 3.4	80.8 ± 13.8	95.0 ± 5.3	101.7 ± 10.9	104.40 ± 19.66
	Medium	29.3 ± 4.8	34.1 ± 7.2	47.9 ± 18.9	29.3 ± 5.5	30.1 ± 16.9	64.9 ± 7.7	80.1 ± 20.6	$\mathbf{95.97 \pm 12.69}$
	M-R	15.4 ± 5.6	20.0 ± 8.4	37.9 ± 12.3	13.6 ± 5.7	5.4 ± 11.0	30.3 ± 2.5	50.4 ± 33.2	$\mathbf{54.12 \pm 15.29}$
	Random	9.8 ± 4.9	24.0 ± 9.8	34.4 ± 5.3	6.3 ± 3.5	-3.8 ± 12.3	6.9 ± 3.1	31.1 ± 6.8	$\mathbf{75.92 \pm 23.81}$
Average rewards		40.7	44.1	58.8	38.3	28.1	49.2	65.8	**82.6**

4.6 Offline MARL Benchmark Results

We evaluate MoSDOT on three offline MARL benchmarks. Tab. 2 reports discrete-action results on SMACv1/v2, and Tab. 3 reports continuous-action results on MPE Simple Spread. On SMAC, generative policies—particularly flow- and diffusion-based methods—tend to perform better than Gaussian and behavior cloning (BC) baselines, consistent with the view that a centralized generative teacher is a useful starting point. On MPE, MoSDOT shows clearer gains on *Medium* and *Random*, where replay contains a wider range of joint actions per observation, and remains comparable to existing generative baselines on *Expert*, where replay is concentrated on a single coordinated mode.

These observations are consistent with the view that capacity-matched source assignment is most useful when replay admits a discernible finite mode support. The effect is naturally smaller when the data concentrates on a single coordinated mode (*Expert*), or when the joint-action space grows faster than per-context replay coverage so that the identity-style aggregation rule used here approaches sample-level coupling (e.g., `terran_10_vs_10`). Designing q_x to recover meaningful mode structure in such regimes is a natural direction for future work.

5 Conclusion

We studied multi-agent generative distillation through how source noise is assigned to coordinated joint behaviors during teacher training. Our analysis shows that endpoint coverage of the teacher is not sufficient for factorized distillation: routing artifacts at the teacher stage propagate to decentralized actors through the distillation loss. MoSDOT addresses this by aligning source noise with a capacity-weighted mode support via conditional SDOT, separating teacher-side routing artifacts from the residual product-projection gap that strict-product execution introduces. A central limitation is the strict-product constraint itself: independent local sampling cannot represent arbitrary correlated joint support, and our shared-randomness variant exposes this gap rather than removing it. Quantifying this residual gap, extending mode-support routing to longer-horizon and value-reweighted settings, and studying when shared execution signals are admissible are natural next steps.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MND) (RS-2026-25553650), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208506), and the InnoCORE program of the Ministry of Science and ICT (N10260099 and N10260110). Prof. Sung-eui Yoon is the corresponding author.

Competing Interests. The authors declare no competing interests.

References

[1] Chao Li, Ziwei Deng, Chenxing Lin, Wenqi Chen, Yongquan Fu, Weiquan Liu, Chenglu Wen, Cheng Wang, and Siqi Shen. DoF: A diffusion factorization framework for offline multi-agent reinforcement learning. In *International Conference on Learning Representations*, pages 34830–34868, 2025.

[2] Dongsu Lee, Daehee Lee, and Amy Zhang. Multi-agent coordination via flow matching. In *International Conference on Learning Representations*, pages 99912–99943, 2026.

[3] Teng Pang, Zhiqiang Dong, Yan Zhang, Rongjian Xu, Guoqiang Wu, and Yilong Yin. Value-guidance MeanFlow for offline multi-agent reinforcement learning. arXiv preprint arXiv:2604.08174, 2026.

[4] Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 48104–48127. PMLR, 2025.

[5] Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. In *Advances in Neural Information Processing Systems*, volume 36, pages 52413–52429, 2023. doi: 10.52202/075280-2282.

[6] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2052–2062. PMLR, 2019.

[7] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.

[8] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20132–20145, 2021.

[9] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=68n2s9ZJWF8`.

[10] Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 10299–10312, 2021.

[11] Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 17221–17237. PMLR, 2022.

[12] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9902–9915. PMLR, 2022.

[13] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=sP1fo2K9DFG`.

[14] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=AHvFDPi-FA`.

[15] Zhengbang Zhu, Minghuan Liu, Liyuan Mao, Bingyi Kang, Minkai Xu, Yong Yu, Stefano Ermon, and Weinan Zhang. MADiff: Offline multi-agent learning with diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pages 4177–4206, 2024. doi: 10.52202/079017-0136.

[16] Ju Dong, Liding Zhang, Lei Zhang, Yu Fu, Kaixin Bai, Zoltán-Csaba Márton, Zhenshan Bing, Zhaopeng Chen, Alois Christian Knoll, and Jianwei Zhang. From flow to one step: Real-time multi-modal trajectory policies via implicit maximum likelihood estimation-based distribution distillation. arXiv preprint arXiv:2603.09415, 2026.

[17] Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. arXiv preprint arXiv:2602.04770, 2026.

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

[19] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=CD9Snc73AW`.

[20] 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*, pages 132826–132844, 2026.

[21] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296, 2017.

[22] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[23] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5887–5896. PMLR, 2019.

[24] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002. doi: 10.1287/moor.27.4.819.297.

[25] Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019. doi: 10.4171/JEMS/889.

[26] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. arXiv preprint arXiv:1902.04043, 2019.

[27] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 37567–37593, 2023. doi: 10.52202/075280-1634.

[28] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pages 6379–6390, 2017.

[29] Krishan Rana, Robert Lee, David Pershouse, and Niko Sünderhauf. IMLE policy: Fast and sample efficient visuomotor policy learning via implicit maximum likelihood estimation. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.158.

[30] Claude Formanek, Asad Jeewa, Jonathan Shock, and Arnu Pretorius. Off-the-grid MARL: Datasets with baselines for offline multi-agent reinforcement learning. arXiv preprint arXiv:2302.00521, 2023.

[31] Jiechuan Jiang and Zongqing Lu. Offline decentralized multi-agent reinforcement learning. In *ECAI 2023: 26th European Conference on Artificial Intelligence*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1148–1155. IOS Press, 2023. doi: 10.3233/FAIA230390.

[32] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv preprint arXiv:1910.00177, 2019.

Appendix Contents

A Proofs and Mathematical Derivations
- A.1 Projection Induced by Local Distillation
- A.2 Block-Locality of Ordinary Conditional Flow Matching
- A.3 Support Error and Reward Ceiling
- A.4 Strict-Product Gap: A Minimal XOR Example
- A.5 Shared Randomness Closes the Strict-Product Gap

B Algorithms
- B.1 MoSDOT (Strict-Product Variant)
- B.2 MoSDOT-SR (Shared-Randomness Variant)
- B.3 Reduction Between the Two Variants

C Limitations and Discussion

D Implementation Details
- D.1 Support Aggregation Rule
- D.2 SDOT Solver and Finite-Cache Rebalancing
- D.3 Teacher Bootstrap for Critic Learning
- D.4 Network Architecture and Flow Schedule

E Experimental Details
- E.1 Benchmark Environments and Datasets
- E.2 Baselines
- E.3 Training and Optimization Protocol
- E.4 Reporting and Statistical Protocol
- E.5 Compute Resources
- E.6 Evaluation Metric Definitions

F Additional Experiments
- F.1 Student-Side Landmark Diagnostics
- F.2 Full Route-Locality Diagnostics
- F.3 Metric Sensitivity Analysis

A Proofs and Mathematical Derivations

This section collects the formal results that justify the design choices of MoSDOT. Appendix A.1 explains why information that is hidden from a local actor is averaged out by ℓ_2 distillation. Appendix A.2 characterizes when conditional flow matching is block-local at the population optimum, and pinpoints the assumptions that MoSDOT deliberately breaks. Appendix A.3 converts off-support fan mass into a value ceiling that even a perfect student cannot escape. Appendix A.4 shows, on the canonical XOR example, that a clean teacher alone cannot remove the strict-product gap, motivating the shared-randomness variant of Section 3. Appendix A.5 then proves the positive direction: shared randomness produces a mixture of product distributions that exactly represents the XOR target, and characterizes the representational range of this family.

A.1 Projection Induced by Local Distillation

This section formalizes the claim that the ℓ_2 distillation loss regresses each local actor onto the conditional mean of the teacher's output given local input. Concretely, we show that the population-optimal local predictor satisfies:

$$f_i^\star(o_i, z_i) \;=\; \mathbb{E}\big[\,[\mu_\psi(\mathbf{x}, z)]_i \,\big|\, o_i, z_i\,\big]\,, \tag{16}$$

so that any component of the teacher's branch decision that remains random after conditioning on (o_i, z_i) is averaged out by the decentralized actor. For any measurable local predictor f, the squared distillation risk decomposes as:

$$\mathbb{E}\big[\|Y - f(S)\|_2^2\big] = \mathbb{E}\Big[\|Y - \mathbb{E}[Y \mid S]\|_2^2\Big] + \mathbb{E}\Big[\|\mathbb{E}[Y \mid S] - f(S)\|_2^2\Big]\,. \tag{17}$$

The first term is independent of f, and the second term is uniquely minimized in L^2 by:

$$f^\star(S) = \mathbb{E}[Y \mid S].$$

Substituting $Y = g_i(O, z_{1:N})$ and $S = (o_i, z_i)$ gives:

$$f_i^\star(o_i, z_i) = \mathbb{E}[g_i(O, z_{1:N}) \mid o_i, z_i]\,.$$

Therefore, any component of the teacher branch decision that remains random after conditioning on the local input is averaged by the decentralized actor. This is the precise sense in which source assignment affects distillation compatibility: a teacher branch whose mode label depends on z_{-i} cannot be recovered by a strict-product local actor that observes only z_i.

A.2 Block-Locality of Ordinary Conditional Flow Matching

This section clarifies when an ordinary conditional flow-matching teacher is block-local at the population optimum. Let

$$X_i = (1 - T_i)Z_i + T_i A_i, \qquad V_i = A_i - Z_i.$$

Assume that, conditioned on the joint context O, the training law factorizes across agents:

$$p(z_{1:N}, a_{1:N}, t_{1:N} \mid O) = \prod_{i=1}^{N} p(z_i, a_i, t_i \mid O).$$

The squared-loss optimal joint flow field is

$$u_i^\star(O, X_{1:N}, T_{1:N}) = \mathbb{E}[V_i \mid O, X_{1:N}, T_{1:N}].$$

Under the conditional factorization above, (V_i, X_i, T_i) is conditionally independent of the off-agent tuples (X_{-i}, T_{-i}) given O, so

$$u_i^\star(O, X_{1:N}, T_{1:N}) = \mathbb{E}[V_i \mid O, X_i, T_i].$$

Although a joint network may have access to off-agent chunks, the population MSE optimum derives no benefit from using them under the factorized training law. This condition fails precisely when the teacher objective introduces coupled endpoint assignments, set-level matching, or other gradients that explicitly depend on full joint branches—which is exactly the regime MoSDOT operates in.

A.3 Support Error and Reward Ceiling

This section formalizes the support-preservation claim used in the main text. For a fixed context x, let the target mode-support measure be

$$\mu^\star_{\mathbf{x}} = \sum_{k=1}^{K} b_k(\mathbf{x})\delta_{y_k(\mathbf{x})}.$$

Let $T_{\mathbf{x}}(z) = y_{\Gamma_{\mathbf{x}}(z)}(\mathbf{x})$ be the action-valued map induced by the source-to-mode assignment. An exact assignment satisfies

$$T_{\mathbf{x}\#}\rho_0 = \sum_{k=1}^{K} \rho_0(\Gamma_{\mathbf{x}}^{-1}(k))\delta_{y_k(\mathbf{x})} = \mu^\star_{\mathbf{x}}.$$

In the ideal assignment limit the source pushforward has zero mass outside the prescribed support and matches the target capacities exactly.

Now let ν be a generated teacher distribution. Let B_k be pairwise disjoint evaluation regions containing the corresponding support elements $y_k(\mathbf{x})$, and let $F = \mathcal{A} \setminus \bigcup_{k=1}^{K} B_k$ denote the remaining off-support fan or bridge region. Define

$$m_k = \nu(B_k), \qquad m_F = \nu(F).$$

The support error on the quotient partition is

$$D_{\mathrm{supp}}(\nu, \mu^\star_{\mathbf{x}}) = \tfrac{1}{2}\left(\sum_{k=1}^{K} |m_k - b_k(\mathbf{x})| + m_F\right),$$

which separates capacity error from off-support mass. If a teacher has small mode-mass error and small fan mass, it is close to the target support distribution under this quotient evaluation.

Assume a bounded reward or value proxy R such that supported modes are optimal and the fan region has a reward gap $\Delta > 0$:

$$R(y_k(\mathbf{x})) = R^\star, \qquad R(a) \le R^\star - \Delta \quad \forall a \in F.$$

Then any teacher distribution ν with fan mass m_F satisfies

$$J_R(\nu) = \mathbb{E}_\nu[R(a)] \le R^\star - \Delta\, m_F.$$

If a student π_S is within total variation distance ϵ of ν, then $\pi_S(F) \ge (m_F - \epsilon)_+$, and therefore

$$J_R(\pi_S) \le R^\star - \Delta(m_F - \epsilon)_+.$$

Hence observed off-support mass imposes a value ceiling on faithful distillation whenever that mass lies in a lower-value region.

A.4 Strict-Product Gap: A Minimal XOR Example

This section illustrates why a clean teacher does not imply exact preservation by a strict-product student. Consider two agents and the target anti-correlated support

$$\mathcal{M} = \{(-1,+1),(+1,-1)\}, \qquad \mu^\star = \tfrac{1}{2}\delta_{(-1,+1)} + \tfrac{1}{2}\delta_{(+1,-1)}.$$

Let the support reward be $R(x,y) = \mathbf{1}[(x,y) \in \mathcal{M}]$. A strict-product student matching the marginals of $\mu^\star$ has

$$q_{\mathrm{clean}}(x,y) = q_x(x)q_y(y), \quad q_x(\pm 1) = q_y(\pm 1) = \tfrac{1}{2},$$

so

$$J_R(q_{\mathrm{clean}}) = q_{\mathrm{clean}}(-1,+1) + q_{\mathrm{clean}}(+1,-1) = \tfrac{1}{2}.$$

Even a clean non-product teacher cannot be exactly represented by independent local sampling.

Now consider a fan-contaminated teacher

$$\nu_m = \tfrac{1-m}{2}\delta_{(-1,+1)} + \tfrac{1-m}{2}\delta_{(+1,-1)} + m\,\delta_{(0,0)}.$$

A strict-product student matching the teacher's marginals independently has $q_x(\pm 1) = q_y(\pm 1) = (1-m)/2$ and $q_x(0) = q_y(0) = m$, with support reward

$$J_R(q_m) = \left(\tfrac{1-m}{2}\right)^2 + \left(\tfrac{1-m}{2}\right)^2 = \tfrac{(1-m)^2}{2}.$$

Therefore

$$J_R(q_{\text{clean}}) - J_R(q_m) = \tfrac{1}{2} - \tfrac{(1-m)^2}{2} = m - \tfrac{m^2}{2} > 0 \quad \text{for } 0 < m \le 1.$$

Removing fan mass does not eliminate the intrinsic product gap, but it strictly improves the marginally distilled product student when fan mass contaminates the local marginals. This is the analytic counterpart of the empirical XOR diagnostic in Fig. 5; the complementary positive result—that the shared-randomness extension (Algorithm 2) does close this gap—is given in Appendix A.5.

A.5 Shared Randomness Closes the Strict-Product Gap

Appendix A.4 shows that a strict-product student of the form Eq. (2) cannot represent the anti-correlated XOR target $\mu^\star = \frac{1}{2}\delta_{(-1,+1)} + \frac{1}{2}\delta_{(+1,-1)}$. This section proves the complementary positive direction for the shared-randomness policy of Eq. (15): a public shared signal h turns the policy into a *mixture of product distributions*, and this mixture exactly represents anti-correlated supports such as XOR. We first derive the mixture form from the one-step implementation, then verify exact XOR reproduction, and finally state the representational range of the resulting policy family.

From one-step implementation to mixture-of-products. Recall the one-step shared-randomness sampler $\hat{a}_i = \mu^{\text{sr}}_{w_i}(o_i, h, \epsilon_i)$ with $h \sim \rho_h$ shared across all agents and $\epsilon_i \sim \rho_i$ drawn independently per agent. We derive Eq. (15) from this implementation in three steps.

(i) Joint conditional law given all noise. Since the sampler is deterministic given $(h, \epsilon_{1:N})$,

$$p(\mathbf{a} \mid \mathbf{x}, h, \epsilon_{1:N}) = \prod_{i=1}^{N} \delta\big(a_i - \mu^{\text{sr}}_{w_i}(o_i, h, \epsilon_i)\big).$$

(ii) Conditional independence given h. The local noises $\{\epsilon_i\}$ are mutually independent, so integrating each ϵ_i against its marginal ρ_i while holding h fixed gives

$$p(\mathbf{a} \mid x, h) = \prod_{i=1}^{N} \underbrace{\int \delta\big(a_i - \mu^{\text{sr}}_{w_i}(o_i, h, \epsilon_i)\big)\, \rho_i(d\epsilon_i)}_{=:\, \pi^{\text{sr}}_{w_i}(a_i \mid o_i, h)} = \prod_{i=1}^{N} \pi^{\text{sr}}_{w_i}(a_i \mid o_i, h).$$

Conditioning on the shared signal restores per-agent independence: the local noises were independent to begin with, and h is the only cross-agent random variable.

(iii) Marginalization over h. Integrating the shared signal yields

$$\pi^{\text{sr}}_w(\mathbf{a} \mid \mathbf{x}) = \int p(\mathbf{a} \mid \mathbf{x}, h)\, \rho_h(dh) = \int \prod_{i=1}^{N} \pi^{\text{sr}}_{w_i}(a_i \mid o_i, h)\, \rho_h(dh),$$

which is exactly Eq. (15). The integrand is a product over agents, but the integral over h produces a *mixture* of such products. Conditional independence given h does not imply marginal independence, so the joint law generally satisfies $a_i \not\perp a_j$ even though $a_i \perp a_j \mid h$.

Exact XOR reproduction. For the target $\mu^\star = \frac{1}{2}\delta_{(-1,+1)} + \frac{1}{2}\delta_{(+1,-1)}$ of Appendix A.4, take a binary shared signal $h \in \{0, 1\}$ with $\rho_h(0) = \rho_h(1) = \frac{1}{2}$ and degenerate local noises ($d_{\epsilon_i} = 0$). Define the per-agent maps

$$\mu^{\text{sr}}_{w_1}(\cdot, h, \cdot) = \begin{cases} -1 & h = 0, \\ +1 & h = 1, \end{cases} \qquad \mu^{\text{sr}}_{w_2}(\cdot, h, \cdot) = \begin{cases} +1 & h = 0, \\ -1 & h = 1. \end{cases}$$

Substituting into Eq. (15) gives

$$\pi^{\text{sr}}_w = \tfrac{1}{2}\,\delta_{-1} \otimes \delta_{+1} + \tfrac{1}{2}\,\delta_{+1} \otimes \delta_{-1} = \mu^\star.$$

Each agent acts deterministically from its own input alone, yet the joint law is exactly the anti-correlated target. The only cross-agent channel is the public draw of h; no observations, actions, or messages are exchanged.

Representational range. The construction generalizes to any target that factorizes through a discrete latent K,

$$\mu^\star(\mathbf{a} \mid \mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \prod_{i=1}^{N} \mu^\star_{i,k}(a_i \mid o_i, \mathbf{x}),$$

by taking ρ_h supported on $\{1, \dots, K\}$ with $\rho_h(k) = \pi_k(\mathbf{x})$ and $\pi^{\mathrm{sr}}_{w_i}(\cdot \mid o_i, k) = \mu^\star_{i,k}(\cdot \mid o_i, \mathbf{x})$. A shared dimension of $\lceil \log_2 K \rceil$ bits suffices for any K-component mixture-of-products target, and richer ρ_h (e.g. $\rho_h = \rho_0$ on $\mathbb{R}^d$) yields a universal class of correlated decentralized policies under independent local computation and a single public coordination channel.

Connection to the MoSDOT-SR algorithm. Algorithm 2 uses the shared signal h as the public latent in this mixture-of-products view. In the Brenier-polar diagnostic instantiation, h is chosen so that the cached SDOT assignment $\Gamma_{\mathbf{x}}(h)$ selects a replay-derived support element, with mixture weights induced by the capacities $b_k(\mathbf{x})$. The generic algorithm does not require exposing the entire teacher source or committing to a particular cell-lookup parameterization; it only requires that all agents receive the same public signal h while retaining local computation.

B Algorithms

This section gives compact pseudocode for the two MoSDOT variants. The mode support, capacities, and SDOT source assignment are already defined in Sec. 3.2–3.3; the algorithms therefore call these steps as subroutines rather than expanding the SDOT dual solver or finite-cache implementation. Aggregation choices, cache rebalancing, and teacher-bootstrap details are deferred to Appendix D.

B.1 MoSDOT (Strict-Product Variant)

Algorithm 1 MoSDOT: SDOT-aligned teacher training and strict-product distillation.

Require: offline replay dataset $\mathcal{D}$; source law ρ_0 over $\mathbf{z} = (z_1, \dots, z_N)$; support aggregation rule $q_{\mathbf{x}}$; cost $c_{\mathbf{x}}$; teacher field u_ψ; decentralized actors $\{\mu_{w_i}\}_{i=1}^N$; centralized critic Q_{tot} and target $\bar{Q}_{\mathrm{tot}}$.
Ensure: decentralized actors $\{\mu_{w_i}\}_{i=1}^N$.

# BUILD ROUTED MODE SUPPORT
for each cached conditioning context $\mathbf{x}$ **do**
  $(\mathcal{Y}_{\mathbf{x}}, b(\mathbf{x})) \leftarrow \textsc{AggregateSupport}(\mathcal{D}, \mathbf{x}; q_{\mathbf{x}})$ ▷ Eqs. (4)–(6)
  $\Gamma_{\mathbf{x}} \leftarrow \textsc{SolveSDOT}(\rho_0, \mathcal{Y}_{\mathbf{x}}, b(\mathbf{x}); c_{\mathbf{x}})$ ▷ Eqs. (8)–(10)
end for
while not converged **do**
  Sample replay batch $\mathcal{B} = \{(\mathbf{x}_b, \mathbf{o}_b, \mathbf{a}_b, r_b, \mathbf{x}'_b, \mathbf{o}'_b, d_b)\}_{b=1}^B \sim \mathcal{D}$.

TRAIN SDOT-ALIGNED TEACHER
  Sample $\mathbf{z}_b \sim \rho_0$ and $t_b \sim \mathrm{Unif}([0,1])$.
  Set routed target $\mathbf{y}_b \leftarrow \mathbf{y}_\Gamma(\mathbf{x}_b, \mathbf{z}_b)$,
  interpolation $\bar{\mathbf{a}}_{b,t} \leftarrow (1 - t_b)\mathbf{z}_b + t_b \mathbf{y}_b$, and
  displacement $\mathbf{v}_b \leftarrow \mathbf{y}_b - \mathbf{z}_b$.
  Update u_ψ using the teacher loss in Eq. (11).
  Define the teacher sampler $\mu_\psi(\mathbf{x}, \mathbf{z}) \leftarrow \mathrm{Flow}_{u_\psi}(\mathbf{x}, \mathbf{z})$.

TRAIN CENTRALIZED CRITIC
  Sample $\mathbf{z}'_b \sim \rho_0$ and bootstrap $\mathbf{a}'_b \leftarrow \mu_\psi(\mathbf{x}'_b, \mathbf{z}'_b)$.
  Update Q_{tot} with the teacher-bootstrap Bellman loss (Appendix D.3) and soft-update $\bar{Q}_{\mathrm{tot}}$.

DISTILL STRICT-PRODUCT ACTORS
  Sample fresh joint source $\boldsymbol{\zeta}_b = (\zeta_{b,1}, \dots, \zeta_{b,N}) \sim \rho_0$.
  Compute $\tilde{\mathbf{a}}_b \leftarrow \mu_\psi(\mathbf{x}_b, \boldsymbol{\zeta}_b)$ and $\hat{\mathbf{a}}_b \leftarrow (\mu_{w_i}(o_{b,i}, \zeta_{b,i}))_{i=1}^N$.
  Update $\{w_i\}_{i=1}^N$ using the value-guided distillation objective in Eq. (14).
end while
return $\{\mu_{w_i}\}_{i=1}^N$.

B.2 MoSDOT-SR (Shared-Randomness Variant)

Algorithm 2 keeps the same mode-support, SDOT, teacher, and critic stages as Algorithm 1. It only changes the actor signature by splitting the teacher source into a public component h and private local noises ϵ_i, matching the shared-randomness policy in Eq. (15).

Algorithm 2 MoSDOT-SR: shared-randomness distillation and execution.

Require: trained MoSDOT teacher μ_ψ and critic Q_{tot}; source factorization $\rho_0 = \rho_h \otimes \bigotimes_{i=1}^N \rho_i$; decentralized actors $\{\mu^{\mathrm{sr}}_{w_i}(o_i, h, \epsilon_i)\}_{i=1}^N$; public seed generator G.
Ensure: decentralized actors $\{\mu^{\mathrm{sr}}_{w_i}\}_{i=1}^N$.

# DISTILL SHARED-RANDOMNESS ACTORS
while not converged **do**
  Sample replay batch $\mathcal{B} \sim \mathcal{D}$, shared signals $h_b \sim \rho_h$, and local noises $\epsilon_{b,i} \sim \rho_i$.
  Set $\mathbf{z}_b = (h_b, \epsilon_{b,1:N})$ and compute $\tilde{\mathbf{a}}_b \leftarrow \mu_\psi(\mathbf{x}_b, \mathbf{z}_b)$.
  Compute student action $\hat{\mathbf{a}}_b \leftarrow (\mu^{\mathrm{sr}}_{w_i}(o_{b,i}, h_b, \epsilon_{b,i}))_{i=1}^N$.
  Update $\{w_i\}_{i=1}^N$ using the same value-guided distillation objective as Algorithm 1.
end while

# DECENTRALIZED EXECUTION
for execution step n **do**
  All agents publicly reconstruct $h_n = G(s_n) \sim \rho_h$; each agent i privately draws $\epsilon_{i,n} \sim \rho_i$.
  Agent i outputs $a_{i,n} = \mu^{\mathrm{sr}}_{w_i}(o_{i,n}, h_n, \epsilon_{i,n})$.
end for
return $\{\mu^{\mathrm{sr}}_{w_i}\}_{i=1}^N$.

B.3 Reduction Between the Two Variants

Algorithm 1 is recovered from Algorithm 2 when no public randomness is used. Equivalently, either the shared component is degenerate ($d_h = 0$) so that $\rho_0 = \bigotimes_i \rho_i$, or the actor ignores h, i.e., $\mu^{\mathrm{sr}}_{w_i}(o_i, h, \epsilon_i) = \mu_{w_i}(o_i, \epsilon_i)$. Then Eq. (15) reduces to the strict-product policy in Eq. (2). At the other endpoint, exposing a non-degenerate public signal makes the deployed law a mixture of product policies, which is the mechanism analyzed in Appendix A.5.

C Limitations and Discussion

We close by collecting the main limitations of MoSDOT and noting partial remedies where they exist.

Strict-product execution is fundamentally limited. Independent local sampling cannot represent arbitrary correlated joint support: even with a perfect teacher, the strict-product student of Eq. (2) fails on anti-correlated targets such as XOR (Appendix A.4). MoSDOT improves the centralized teacher but does not remove this execution-time constraint. The shared-randomness variant (Appendix A.5) closes the gap on supports representable as mixtures of products, but requires a public coordination channel, which is not always admissible in deployment. Quantifying when such a channel is available—and how much shared dimension is actually needed—remains open.

Empirical scope. We evaluate on three offline MARL benchmarks (SMACv1, SMACv2, MPE Simple Spread) with at most 10 agents and either small discrete or low-dimensional continuous action spaces. Scaling to larger agent populations, richer action spaces, or online fine-tuning is left for future work, as is direct comparison with communication-augmented baselines that use richer cross-agent channels than the public seed considered here.

D Implementation Details

This section documents implementation choices that are specific to MoSDOT itself, as opposed to general experimental settings that apply to all methods (those are deferred to Appendix E).

D.1 Support Aggregation Rule

The aggregation rule $q_{\mathbf{x}}$ specifies the granularity of the replay-derived target support and is not treated as an oracle mode label. The identity rule $q_{\mathbf{x}}(m) = m$ is always admissible and recovers the raw sample support. When repeated, quantized, or high-confidence clustered joint targets are present, $q_{\mathbf{x}}$ may merge those targets into a smaller support. We use *conservative aggregation*: ambiguous targets remain separate rather than being forced into a shared support element. The representation used by $q_{\mathbf{x}}$ can be a joint action, an action chunk, a short-horizon endpoint, or a learned target embedding. In all cases the support index refers to a full joint target rather than independently chosen per-agent modes. This convention prevents the assignment problem from degenerating into a product of local mode assignments. For discrete-action benchmarks we use the identity rule; for continuous-action benchmarks we use a simple joint-action quantization at fixed bin width.

D.2 SDOT Solver and Finite-Cache Rebalancing

The Laguerre assignment Γ_x defined in Sec. 3 is specified for the population source law ρ_0. In implementation we approximate ρ_0 by a finite Monte-Carlo cache: for each context/support group c we draw $\mathcal{Z}_c = \{z^\ell\}_{\ell=1}^{L} \overset{\text{i.i.d.}}{\sim} \rho_0$ of size L (65,536 in the landmark diagnostic) and define the empirical cell masses

$$\hat{m}_{c,k}^{(L)} = \tfrac{1}{L}\,\big|\{\ell : \Gamma_c(z^\ell) = k\}\big|.$$

The cost c_c is the squared Euclidean distance in the joint-action representation that defines $\mathcal{Y}_c$.

Dual solver. The SDOT dual is a K_c-dimensional concave program. We estimate the dual weights using stochastic gradient updates with a fixed iteration budget, then evaluate the resulting assignment on a separately sampled finite Gaussian cache.

Finite-cache rebalancing. With finite L, residual deviations between $\hat{m}_{c,k}^{(L)}$ and $b_{c,k}$ may persist after dual optimization. Following the rebalancing principle of Kong et al. [20], we preserve as many original source–target assignments as possible while adjusting target counts to match the prescribed frequencies up to integer rounding. For uniform target weights, the rebalanced counts differ by at most one. This step acts only on the finite cache and does not modify the population SDOT problem; in our landmark caches it reassigns 0.6% to 1.6% of cached entries.

D.3 Teacher Bootstrap for Critic Learning

In our value-guided implementation, the assignment-trained teacher provides bootstrap actions for critic learning. For a replay transition (o_i, a_i, r_i, o_i') we sample a joint next action from the teacher,

$$\mathbf{a}' = \mu_\psi(o', z'), \qquad z' \sim \rho_0,$$

and train an individual critic with

$$\mathcal{L}_Q(\theta_i) = \mathbb{E}\left[\Big(Q_{\theta_i}(o_i, a_i) - \big(r_i + \gamma(1-d)\bar{Q}_{\bar{\theta}_i}(o_i', a_i')\big)\Big)^2\right].$$

The decentralized actors are then updated with the value-guided distillation objective in Eq. (14). This teacher-bootstrap variant is an implementation choice; the core MoSDOT construction is the source assignment used to train the centralized teacher.

D.4 Network Architecture and Flow Schedule

All networks are $[512, 512, 512, 512]$ MLPs with layer normalization. The teacher conditions on the joint observation and joint noise; the decentralized students take only their local observation and local noise slice. The MoSDOT-SR variant additionally uses a trainable noise map $r_{\eta_i}(o_i, h, \epsilon_i)$ that combines the public shared signal h with the local noise ϵ_i. The teacher uses a linear-interpolation flow with 10 Euler steps for both training-time bootstrap actions and inference.

E Experimental Details

E.1 Benchmark Environments and Datasets

We evaluate on three offline MARL benchmarks. The choice of maps, quality regimes, and dataset sources follows the conventions of Lee et al. [2] and Formanek et al. [30] so that comparator scores remain directly comparable.

SMACv1. SMACv1 [26] simulates decentralized StarCraft II combat between a built-in AI and learned policies. Agents observe partial local information within a sight range, and the discrete action space includes movement, target attack, stop, and no-op. We evaluate on five maps spanning homogeneous and heterogeneous unit compositions: 3m, 8m, 2s3z, 5m_vs_6m, and 2c_vs_64zg. For each map we use the Good, Medium, and Poor quality datasets from the off-the-grid benchmark [30], giving 15 dataset–scenario pairs.

SMACv2. SMACv2 [27] extends SMACv1 with randomized start positions, randomized unit types, and rebalanced sight and attack ranges, increasing the diversity of episodes. We evaluate on three maps under the Replay regime: terran_5_vs_5, zerg_5_vs_5, and terran_10_vs_10. Datasets are again drawn from the off-the-grid benchmark.

MPE Simple Spread. MPE [28] is a continuous-action particle environment for cooperative coordination. In Simple Spread, agents must cover landmarks while avoiding collisions. We use the four quality regimes (Expert, Medium, Medium-Replay, Random) collected by Pan et al. [11]. Following Lee et al. [2], returns are normalized as

$$100 \times \frac{S - S_{\text{random}}}{S_{\text{expert}} - S_{\text{random}}},$$

with $S_{\text{random}} = 159.8$ and $S_{\text{expert}} = 516.8$.

E.2 Baselines

Diagnostic baselines. On the landmark and XOR diagnostics, we compare against three teacher parameterizations. *Independent flow matching* [18] pairs source noise and replay targets independently within each context. *IMLE policy* [29] uses a set-level objective based on implicit maximum likelihood estimation to improve endpoint coverage and reduce mode collapse. *Drifting* [17] trains generated samples through attraction toward the target support and repulsion among generated samples.

Benchmark baselines. For benchmark evaluation we additionally compare against three categories of offline MARL baselines reported in Lee et al. [2], all reproduced under the same dataset splits and seed protocol.

Gaussian-policy baselines. BC is a behavioral cloning policy parameterized as a Gaussian (continuous) or categorical (discrete) head. MABCQ [31] extends BCQ [6] to multi-agent settings with a Q-mixer. MACQL extends CQL [7] with a conservative penalty. MATD3BC extends TD3+BC [8] with $\alpha = 2.5$. ICQ [10] uses an implicit constraint approach for joint policy regularization. OMAR [11] combines CQL-style critic regularization with a CEM-based policy improvement head. OMIGA [5] uses a state-dependent value mixer with advantage-weighted regression [32].

Diffusion-policy baselines. Diffusion BC is a behavioral cloning policy with a 200-step Gaussian DDPM scheduler. MADiff [15] models coordinated multi-agent trajectories with a centralized-then-decentralized diffusion policy. DoF [1] factorizes the diffusion process across agents under the IGM-compatible W-concat factorization with 200 denoising steps.

Flow-policy baselines. Flow BC trains an independent flow policy per agent with the same 10-step linear-interpolation schedule as MAC-Flow but without the joint teacher and IGM critic. MAC-Flow [2] is the immediate predecessor of MoSDOT and shares the centralized-teacher to decentralized one-step actor pipeline. MoSDOT replaces the independent source–target pairing of MAC-Flow's teacher with the SDOT-aligned coupling.

All baselines use the same $[512, 512, 512, 512]$ MLP architecture as MoSDOT and follow the official implementations cited above. Public repositories used for reimplementation are listed in the supplementary code release.

E.3 Training and Optimization Protocol

Optimization. We use Adam with learning rate 3×10^{-4} for both actor and critic, batch size 64, weight decay 0, optimizer epsilon 10^{-5}, Polyak averaging coefficient $\tau = 0.005$, and discount factor $\gamma = 0.995$. The BC distillation coefficient is $\alpha = 3.0$ by default.

Training schedule. We train for 10^6 gradient steps on SMACv1 and SMACv2, and 5×10^5 steps on MPE, matching the budget of Lee et al. [2]. We evaluate every 50k steps with 10 evaluation episodes.

E.4 Reporting and Statistical Protocol

For benchmark scores we run 6 random seeds and report mean $\pm$ two standard deviations following Lee et al. [2]. Unless otherwise specified, diagnostic $\pm$ values denote one standard deviation across 4 seeds. Tables 7 and 8 use seed–agent aggregation for the metrics identified in their captions. Best and second-best values are highlighted according to their numerical ranking, where indicated.

In Fig. 4 (left), bars show mean joint mode coverage and error bars indicate one standard deviation across four seeds; hatched and solid bars denote teachers and students, respectively. The right panel shows teacher Jacobian ratios, with one agent-averaged value per seed. Boxes span the interquartile range, central lines mark medians, and whiskers extend to the most extreme observations within 1.5 interquartile ranges of the box; points show the individual seed values.

E.5 Compute Resources

All experiments run on a single NVIDIA RTX 4090 GPU with CUDA 12.2 and JAX 0.4.30. SDOT assignments are precomputed before teacher training and reused throughout optimization.

Preprocessing and training time. Table 4 compares standard flow matching and MoSDOT on the landmark diagnostic under matched data, architecture, minibatch size, and training budget. Both methods require dataset preparation; MoSDOT additionally performs dual optimization, Laguerre assignment, finite-cache rebalancing, and cache storage. The additional preprocessing is paid once, and the deployed actor remains the same one-step network.

Method	Preprocessing (s)	Training (min)	Total (min)
Standard flow matching	70	37.75 ± 0.01	38.9
MoSDOT	279	38.88 ± 0.29	43.5

Table 4: **Landmark wall-clock comparison.** Training uses $40{,}000$ steps; training times are mean $\pm$ one standard deviation over four seeds. Total time includes one preprocessing pass and the mean training time.

Occupied support and cache construction. Table 5 reports retained replay rows M, occupied support elements K, and one-time cache-construction times for the measured benchmark configurations. Support size is bounded by the retained replay data ($K \leq M$), rather than the full Cartesian joint-action space. It need not increase monotonically with the number of agents: replay diversity and aggregation granularity also affect K.

Agents	Dataset	Retained rows M	Occupied support K	K/M	Cache time (s)
3	SMACv1 3m / Poor	200,000	6,625	3.3%	10.5
3	MPE Spread / Medium	5,000,000	72,509	1.5%	122.4
5	SMACv1 5m_vs_6m / Good	200,000	20,454	10.2%	10.6
5	SMACv2 zerg_5_vs_5 / Replay	200,000	2,380	1.2%	38.6
8	SMACv1 8m / Poor	200,000	32,653	16.3%	67.1
10	SMACv2 terran_10_vs_10 / Replay	200,000	88,729	44.4%	147.7

Table 5: **Occupied replay support and one-time cache cost.** M and K are totals over the retained dataset, not counts of all possible joint actions. Times measure cache construction before training.

Assignment cost versus the complete dual solve. For an SDOT problem with K occupied support elements in a d-dimensional joint-action representation, assigning L source samples by exhaustive cost comparison takes $O(LKd)$ operations. With T stochastic dual updates using B source samples per update, dual optimization requires $O(TBKd)$ distance-evaluation work; assigning the final cache adds $O(LKd)$. Thus the dominant distance-evaluation cost is

$$O((TB + L)Kd)\,.$$

For multiple conditioning groups, these costs add across groups. The measured cache times also include preprocessing, rebalancing, and storage. With fixed per-agent action dimension, d grows linearly with the number of agents, while K depends on the occupied replay support. Neither the dual solve nor cache construction is repeated during decentralized execution.

E.6 Evaluation Metric Definitions

This subsection defines the diagnostic metrics used in Section 4 and Appendix F. Teacher-side metrics (Appendix E.6.1) test whether the teacher reaches the mode support, and route-locality metrics (Appendix E.6.2) test whether mode choices depend only on locally observable information.

E.6.1 Teacher-Side Support Metrics

Fan-free. For the landmark diagnostic, an agent's initial action is valid if its direction lies within 10° of either of its two landmark directions; actions with norm below 10^{-6} are invalid. Valid actions receive the nearest-direction branch label. *Fan-free* is the valid-action fraction, averaged over agents. Higher is better.

Balance. Let $n_{i,0}$ and $n_{i,1}$ count valid actions assigned to agent i's two local modes. *Balance* reports the minimum across agents of

$$\frac{2\min(n_{i,0}, n_{i,1})}{n_{i,0} + n_{i,1}},$$

with value 0 if no action is valid. Higher is better.

Success. *Success* is the fraction of agent trajectories that come within distance 0.05 of either of that agent's landmarks at any time during the five-step rollout, averaged over agents and rollouts.

Route consistency. Route consistency is the product of the fan-free fraction and one minus the agent-averaged KNN flip rate at $k = 20$, defined below. Thus it rewards both valid actions and stable local branch labels.

Joint mode coverage. Let V denote the event that all agents' actions are valid, and let $\mathbf{m} = (m_1, \ldots, m_N)$ be their joint branch-label tuple. Joint mode coverage is $\Pr(V)\, H(\mathbf{m} \mid V)/\log_2(2^N)$, with entropy measured in bits and value 0 when no joint sample is valid.

E.6.2 Route-Locality Metrics

Route locality is reported through two complementary families. The sample-level family operates on replayed teacher samples. The model-level family operates on the restored teacher MLP and provides causal evidence for cross-noise dependence.

Own-cell entropy. For each agent i, samples with valid branch labels $m_i \in \{0, 1\}$ are retained, and their z_i values are partitioned into a $b \times b$ quantile grid with default $b = 10$. We report the cell-mass-weighted average of empirical branch entropy in bits, normalized over retained cells. Cells require at least 20 samples in Table 7 and 4 in the sensitivity analysis of Table 8. The own-cell mix rate uses the same cells and weights, replacing entropy by $\min(p, 1 - p)$, where p is the fraction of branch-1 labels within a cell.

KNN flip rate. Among samples with valid branch labels, we find each sample's k nearest neighbors in z_i, excluding itself, with default $k = 20$. We report the mean fraction of neighbors with a different branch label m_i. Sensitivity at $k \in \{10, 20, 50\}$ is reported in Appendix F.3.

Off-diagonal AUC excess. For each $j \neq i$, we fit a Fisher linear discriminant to predict valid branch labels m_i from z_j and evaluate its AUC on those samples. We report $\max_{j \neq i} \max(\mathrm{AUC}_{ij}, 1 - \mathrm{AUC}_{ij}) - 0.5$. A value near zero indicates little linearly decodable cross-noise dependence.

Jacobian off-diagonal ratio. On the restored teacher, we differentiate the full endpoint sampler with respect to source noise via forward-mode autodiff. Writing $J_{ij}^{(s)} = \partial a_i^{(s)} / \partial z_j^{(s)}$ for sample s, the reported ratio is

$$\frac{\mathbb{E}_{s,i,j \neq i} \|J_{ij}^{(s)}\|_F}{\mathbb{E}_{s,i} \|J_{ii}^{(s)}\|_F},$$

where each expectation is a uniform empirical average, with a numerical floor on the denominator. The ratio uses the unclipped endpoint helper so that action saturation does not mask cross-noise sensitivity. The student ratio is exactly 0 by architecture and serves as a structural reference.

Counterfactual flip rate. We fix z_i, resample z_{-i} from ρ_0 32 times per sample, and pass each through the restored teacher sampler. Among pairs with valid branch labels for both the original and resampled output, the flip rate is the fraction that change agent i's branch label. This is the causal counterpart of the KNN flip diagnostic. It tests whether the model itself uses z_{-i} to select agent i's branch, rather than whether the population of trained samples happens to mix.

F Additional Experiments

This section collects supplementary numerical results referenced from the main text. Appendices F.1–F.3 report extended diagnostic results on the controlled tasks: student-side endpoint quality, full sample- and model-level route-locality (entanglement) diagnostics, and a sensitivity study showing that the route-locality ranking is invariant to bin and neighborhood choices.

F.1 Student-Side Landmark Diagnostics

Tab. 6 reports endpoint-quality metrics for one-step factorized students distilled from each teacher under the same student architecture, seed list, and training budget. Comparing this table with the corresponding teacher-side numbers in Tab. 1 quantifies the teacher-to-student transfer described in Section 4: the MoSDOT student largely retains the teacher's fan-free, success, and balance scores, whereas IMLE and Drifting students lose substantial endpoint quality through distillation. This is consistent with the projection result of Appendix A.1—only teachers whose mode choice is locally identifiable can be reproduced by a strict-product student. Figure 7 shows the corresponding joint mode-tuple distributions.

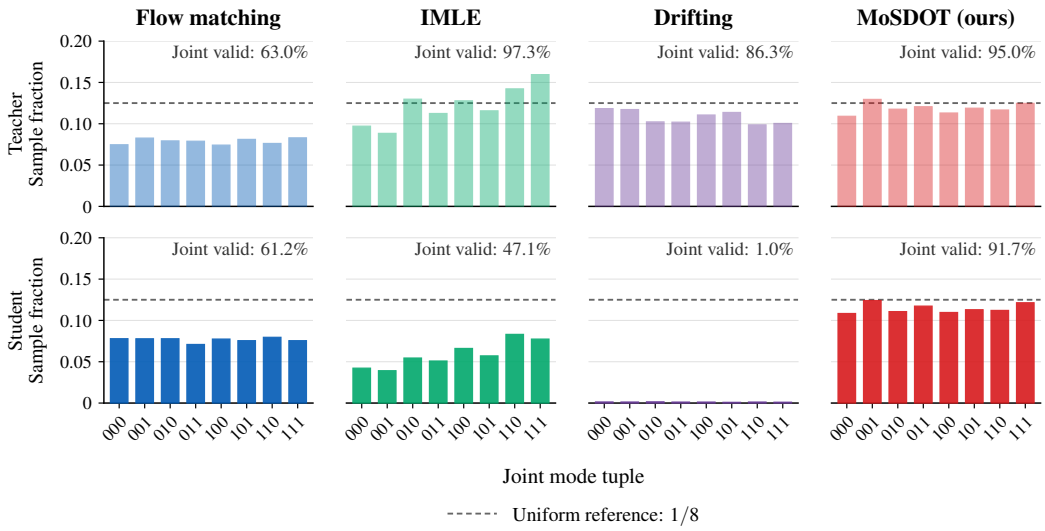

Figure 7: **Joint mode-tuple support, teacher vs. student.** Bars show the fraction of 32,768 samples (fan included in the denominator) falling into each of the 8 joint mode tuples on the 3-agent landmark task; the dashed line marks the zero-fan uniform ideal $1/8$. Top row: centralized teacher (joint flow ODE). Bottom row: one-step factorized student distilled from each teacher under matched architecture and budget. MoSDOT preserves both validity and balance through distillation (95.0% $\rightarrow$ 91.7% valid), while IMLE loses half its valid mass (97.3% $\rightarrow$ 47.1%) and Drift collapses entirely to the fan region (86.3% $\rightarrow$ 1.0%). This is the per-mode visualization of the aggregate scores in Tab. 6 and the empirical counterpart of the projection result in Appendix A.1.

Method	Fan-free ↑	Success ↑	Balance ↑
Flow matching	0.849 ±0.004	0.846 ±0.038	0.949 ±0.024
IMLE	0.784 ±0.011	0.558 ±0.072	0.621 ±0.104
Drifting	0.364 ±0.078	0.685 ±0.071	0.328 ±0.417
MoSDOT (ours)	**0.971** ±0.003	**0.971** ±0.004	**0.963** ±0.019

Table 6: **One-step student landmark diagnostics.** Mean ± one standard deviation over 4 seeds; same 8192 noises × 3 agents per seed as the teacher table. Best and second-best in sage and powder blue. Higher is better for all metrics.

F.2 Full Route-Locality Diagnostics

Table 7 reports the full route-locality diagnostics, combining sample-level and MLP-level measurements. The sample-level columns (own-cell entropy, mix rate, KNN flip, off-diagonal AUC excess) are computed on replayed teacher samples and capture whether nearby source points within an agent's noise slice receive consistent mode labels. The MLP-level columns (endpoint Jacobian ratio, counterfactual flip rate) are computed by re-evaluating the restored teacher MLP under controlled noise perturbations, providing *causal* evidence that the model itself uses z_{-i} to choose agent i's branch rather than this dependence being an artifact of training-sample correlations.
Across both families, MoSDOT is closest to the lowest-entanglement reference (independent flow matching) and substantially below IMLE and Drifting, while simultaneously producing the cleanest endpoints (Table 1). This is the empirical counterpart of Appendix A.2. When the teacher objective explicitly couples joint endpoints (IMLE, Drifting), block-locality is broken and cross-noise dependence appears. When the coupling is performed upstream by SDOT, endpoint quality improves *without* reintroducing cross-noise dependence.

Method	*Sample-level*				*MLP-level*	
	$H(m_i \mid z_i \text{ cell})$	Mix rate	KNN flip	AUC excess	Jac. ratio	CF flip
	(lower is better)				(lower is better)	
Flow	0.019 $\pm$0.024	0.006 $\pm$0.010	0.001 $\pm$0.000	0.020 $\pm$0.008	0.066 $\pm$0.005	0.000 $\pm$0.000
IMLE	0.319 $\pm$0.109	0.098 $\pm$0.036	0.135 $\pm$0.051	0.105 $\pm$0.042	0.292 $\pm$0.055	0.129 $\pm$0.050
Drift	0.666 $\pm$0.252	0.244 $\pm$0.117	0.315 $\pm$0.134	0.290 $\pm$0.123	0.772 $\pm$0.148	0.307 $\pm$0.128
Ours	0.064 $\pm$0.024	0.020 $\pm$0.010	0.011 $\pm$0.003	0.012 $\pm$0.003	0.046 $\pm$0.015	0.002 $\pm$0.002

Table 7: Full route-locality diagnostics. Sample-level metrics use 12 points per method (4 seeds $\times$ 3 agents). H denotes own-noise cell entropy with 10 bins, Mix denotes own-cell mix rate, KNN flip uses k=20, and AUC excess reports the maximum off-diagonal LDA AUC minus 0.5. MLP-level metrics report 4 points for the Jacobian ratio (4 seeds, agent-averaged) and 12 points for CF flip (4 seeds $\times$ 3 agents). All values are mean $\pm$ std.

F.3 Metric Sensitivity Analysis

Table 8 reports own-cell entropy and KNN flip across bin counts $b \in \{5, 10, 20\}$ and neighborhood sizes $k \in \{10, 20, 50\}$ for the four diagnostic comparators. The relative ordering Flow $\leq$ Ours $\ll$ IMLE $<$ Drift is preserved across all hyperparameter settings, ruling out the hypothesis that the route-locality result depends on a single grid choice.

Method	H, b=5	H, b=10	H, b=20	KNN k=10	KNN k=20	KNN k=50
Flow	0.109 $\pm$0.061	0.021 $\pm$0.024	0.005 $\pm$0.006	0.000 $\pm$0.000	0.001 $\pm$0.000	0.001 $\pm$0.001
IMLE	0.360 $\pm$0.101	0.319 $\pm$0.109	0.291 $\pm$0.106	0.134 $\pm$0.051	0.135 $\pm$0.051	0.137 $\pm$0.051
Drift	0.690 $\pm$0.240	0.666 $\pm$0.252	0.634 $\pm$0.252	0.314 $\pm$0.134	0.315 $\pm$0.134	0.316 $\pm$0.133
Ours	0.181 $\pm$0.015	0.064 $\pm$0.024	0.032 $\pm$0.010	0.008 $\pm$0.003	0.011 $\pm$0.003	0.019 $\pm$0.003

Table 8: Route-locality sensitivity. Each entry reports 12 points per method (4 seeds $\times$ 3 agents), mean $\pm$ std. The relative ranking is preserved across all bin counts and neighborhood sizes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions (MoSDOT teacher-side source assignment, the strict-product gap, and the shared-randomness variant), which are supported by the diagnostics in Sec. 4 and the benchmarks in Sec. 4.6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 4 discusses regimes where the effect of MoSDOT diminishes, and we explicitly states the strict-product constraint and remaining open directions as limitations in Appendix C.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All formal results (projection induced by local distillation, block-locality of conditional flow matching, support error and reward ceiling, and the XOR strict-product gap) are stated and proved in Appendix A.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithms 1 and 2 give full pseudocode, and Appendices D and E specify the architecture, flow schedule, SDOT solver, optimizer, training budget, datasets, and evaluation protocol. The code will also be made publicly available.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not released at submission time. We use publicly available benchmarks (SMACv1, SMACv2, MPE) with the off-the-grid datasets cited in Appendix E.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E specifies the optimizer, learning rate, batch size, training schedule, and evaluation protocol.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Benchmark scores report mean $\pm 2\sigma$ over 6 seeds. Diagnostic aggregation units and uncertainty summaries are specified in Appendix E.4 and the corresponding captions.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Mentioned in Appendix E.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses standard simulation benchmarks with no human subjects, personal data, or sensitive content.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational research on offline multi-agent RL methodology evaluated on simulation benchmarks, with no direct path to specific societal applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, generative content, or scraped data that pose misuse risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All benchmarks (SMACv1, SMACv2, MPE) and baselines are cited in Sec. 4 and Appendix E, and used under their original licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or pre-trained models at submission time.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A]

Justification: LLMs are not part of the core methodology; the proposed method, experiments, and analysis do not involve LLMs.