MEAN-FIELD DIFFUSER: SCALING OFFLINE MARL TO THOUSANDS OF AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based planning has achieved strong results in single-agent offline reinforcement learning, yet scaling to many-agent systems remains intractable due to the curse of dimensionality in the joint trajectory space. We introduce MF-Diffuser, a framework that lifts trajectory planning to the Wasserstein space of trajectory distributions, where the propagation of chaos ensures a small representative subset of agents captures the full population dynamics. Our approach features a *value-weighted chaotic entropy* objective that reconciles generative fidelity with return maximization, and a hierarchical coarse-to-fine strategy that progressively grows the agent population during denoising. We establish end-to-end suboptimality bounds with four interpretable terms, revealing that mean-field approximation error scales as $\mathcal{O}(H^2/\sqrt{N})$ while the offline distribution-shift term is independent of population size N, and prove the generated policy is an approximate mean-field Nash equilibrium with explicit convergence guarantees. Experiments on three mean-field RL benchmarks—spanning stage games, sequential dynamics, and adversarial team competition—show MF-Diffuser achieves the best return in the majority of settings, with the largest gains on suboptimal offline data and at extreme scales ($N \geq 10^3$).

1 INTRODUCTION

Offline reinforcement learning (offline RL) (Levine et al., 2020; Lange et al., 2012) has emerged as a principled framework for learning decision-making policies from pre-collected datasets without further environment interaction. Among recent advances, diffusion-based planning methods such as Diffuser (Janner et al., 2022) and Decision Diffuser (Ajay et al., 2023) have demonstrated remarkable effectiveness by casting trajectory optimization as a conditional generative modeling problem: a denoising diffusion model is trained to capture the distribution of trajectories in the offline dataset, and high-return plans are generated at inference time via guided sampling.

Despite their empirical success, these diffusion-based planners are inherently designed for **single-agent** settings. Extending them to **many-agent systems**—environments populated by hundreds or thousands of interacting agents—poses a fundamental scalability challenge. Consider N homogeneous agents, each with state space $\mathcal{S}$ and action space $\mathcal{A}$. A joint trajectory over horizon H lives in $\mathbb{R}^{N\times(d_s+d_a)\times H}$. Even for moderate values of N, the dimensionality of this joint space grows linearly in N, leading to an exponential blowup in the sample complexity required for accurate diffusion modeling (De Bortoli, 2022; Chen et al., 2023). This is precisely the *curse of dimensionality* that plagues conventional approaches.

Meanwhile, many-agent decision-making problems are ubiquitous in real-world applications. Large-scale traffic control (Vinitsky et al., 2018), swarm robotics (Dorigo et al., 2021), financial market modeling (Lachapelle & Wolfram, 2011), and epidemiological policy design (Elie et al., 2020) all involve populations of agents whose individual decisions are coupled through shared environmental dynamics. For such problems, **mean-field reinforcement learning** (MFRL) (Yang et al., 2018; Gu et al., 2025) provides a principled approximation: by replacing the explicit dependence on all other agents with a dependence on the *population distribution* (mean field), the complexity of the N-agent problem is reduced to that of a representative single agent interacting with the population.

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

In this paper, we introduce **MF-Diffuser** (**M**ean-**F**ield Diffuser), a diffusion-based planning framework that bridges the gap between trajectory-level generative planning and mean-field multi-agent systems. Our key insight is to model the N-agent trajectory planning problem through the lens of *interacting particle systems*: by treating each agent's trajectory as a particle in a high-dimensional trajectory space, the joint planning problem becomes amenable to mean-field analysis, where the propagation of chaos property (Sznitman, 1991) ensures that the behavior of the full population can be approximated by a representative subset.

Transferring mean-field particle theory to the sequential decision-making domain introduces three challenges absent from static settings: **(C1)** agent trajectories carry *temporal coupling*—interaction effects accumulate over the planning horizon H, causing approximation errors to compound multiplicatively; **(C2)** the objective combines distributional matching with *return maximization*; and **(C3)** offline data introduces *distribution shift* whose interaction with the mean-field structure must be controlled.

MF-Diffuser addresses all three challenges with the following contributions: (1) **Mean-field Trajectory Diffusion with Value-weighted Chaotic Entropy** (C1, C2, C3). We formulate many-agent offline RL as generative modeling on the Wasserstein space of trajectory distributions, defining mean-field SDEs over trajectories with temporally-structured interaction kernels. Building on the RL-as-inference framework (Levine, 2018), we derive a mean-field value score matching (MF-VSM) objective that unifies distributional fidelity with return maximization at the chaotic limit. (2) **Hierarchical Coarse-to-Fine Planning** (C1). Justified by propagation of chaos ($M = \tilde{\mathcal{O}}(\sqrt{N})$ agents suffice), we progressively grow the agent population during denoising—starting from a small representative group and branching to the full population. (3) **End-to-end Planning Guarantee** (C1, C3). Theorem 4.2 decomposes the suboptimality into four interpretable terms: score matching error, subdivision error with geometric decay, mean-field approximation scaling as $\mathcal{O}(H^2/\sqrt{N})$, and an offline distribution-shift term independent of N. (4) **Game-theoretic Guarantees** (C2, C3). Theorem 4.6 shows the generated policy is an approximate mean-field Nash equilibrium with $\mathcal{O}(1/\sqrt{N})$ rate; under Lasry–Lions monotonicity we further establish convergence to the unique MFE (Theorem D.13) and characterize the social welfare–Nash efficiency gap (Proposition D.15; both in Appendix D.4).

2 Problem Setup

Mean-field MDP. We consider a symmetric N-agent MDP. At step $h \in \{0, \dots, H-1\}$, agent i has state $\mathbf{s}_h^i \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$, action $\mathbf{a}_h^i \in \mathcal{A} \subseteq \mathbb{R}^{d_a}$, with discount $\gamma \in (0,1)$. The **mean-field interaction** assumption stipulates that dynamics and rewards depend on others only through the *empirical state distribution* $\bar{\mu}_h^N = \frac{1}{N}\sum_{j=1}^N \delta_{\mathbf{s}_h^j}$:

$$\mathbf{s}_{h+1}^i \sim P(\cdot|\mathbf{s}_h^i, \mathbf{a}_h^i, \bar{\mu}_h^N), \qquad r_h^i = r(\mathbf{s}_h^i, \mathbf{a}_h^i, \bar{\mu}_h^N). \tag{1}$$

Agents are **homogeneous** (sharing P and r); we seek a common policy $\pi : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{A})$ maximizing the social welfare

$$\max_\pi J(\pi) = \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h \frac{1}{N}\sum_{i=1}^N r(\mathbf{s}_h^i, \mathbf{a}_h^i, \bar{\mu}_h^N)\right]. \tag{2}$$

As $N \to \infty$, $\bar{\mu}_h^N$ converges to a deterministic flow μ_h, reducing the problem to a representative agent. The **offline dataset** $\mathcal{D}$ consists of N-agent episodes collected under behavior policy π_β; trajectories $\boldsymbol{\tau}^i = (\mathbf{s}_0^i, \mathbf{a}_0^i, \dots, \mathbf{s}_H^i) \in \mathbb{R}^{D_\tau}$ with $D_\tau = (d_s + d_a)H + d_s$. Mean-field Q-learning (MFQ) (Yang et al., 2018) factors the joint Q-function as $Q^{MF}(\mathbf{s}^i, \mathbf{a}^i, \bar{\mu}_\mathbf{a})$ via mean-action coupling and serves as our behavior-policy collector for the offline data.

Score-based Diffusion Trajectory Planning. Diffuser (Janner et al., 2022) models the trajectory distribution $p(\boldsymbol{\tau})$ in $\mathcal{D}$ via a forward SDE that progressively corrupts $\boldsymbol{\tau}$ with Gaussian noise, $d\boldsymbol{\tau}_u = f_u(\boldsymbol{\tau}_u)du + \sigma_u dB_u$ on $u \in [0, T]$, and a learned reverse SDE that recovers samples by following the score $\nabla \log \zeta_t(\boldsymbol{\tau}_t)$ of the noised marginal. The score network $\mathbf{s}_\theta$ is trained by score matching, $\mathcal{J}_{SM}(\theta) = \mathbb{E}_{t,\boldsymbol{\tau}_t}\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t) - \nabla \log \zeta_t(\boldsymbol{\tau}_t)\|^2$. At inference, trajectories are generated by simulating the reverse SDE from Gaussian noise; high-return plans are obtained by adding a value-gradient

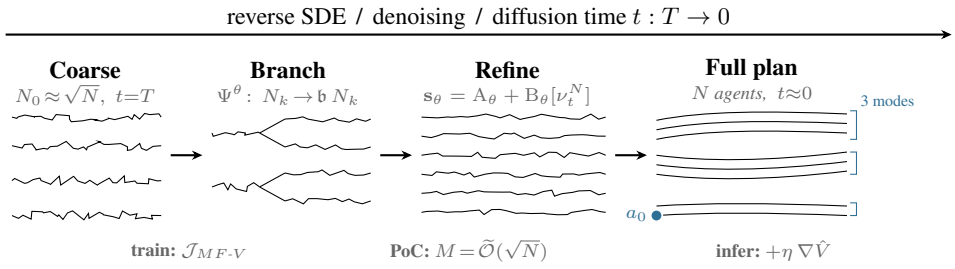


Figure 1: **MF-Diffuser method overview.** Inference proceeds left-to-right along reverse diffusion time $t : T \to 0$. (*Coarse*) start from $N_0 \approx \sqrt{N}$ Gaussian-noise trajectories; (*Branch*) the operator $\Psi^\theta : N_k \to \mathfrak{b} N_k$ spawns child trajectories at designated denoising steps; (*Refine*) the mean-field score $\mathbf{s}_\theta = \mathrm{A}_\theta + \mathrm{B}_\theta[\nu_t^N]$ couples individual dynamics with the population through B_θ; (*Full plan*) at $t \approx 0$ the N trajectories concentrate on the data manifold (e.g. 3 modes shown), and only the first action a_0 of each is executed before re-planning.

bias $\eta\, \nabla_{\boldsymbol{\tau}} \hat{V}(\boldsymbol{\tau})$ to the score, and the observed initial state is enforced via inpainting. Decision Diffuser (Ajay et al., 2023) replaces the gradient bias with classifier-free return-bucket conditioning. We build on this single-agent paradigm but lift it to the joint N-agent setting via mean-field analysis. An extended review is in Appendix B.

3 Mean-field Diffuser

MF-Diffuser recasts many-agent offline RL as mean-field trajectory generation; Figure 1 sketches the inference pipeline. The three components below address the curse of dimensionality.

3.1 Mean-field Trajectory SDEs

We model each agent's trajectory as a particle in $\mathcal{X} := \mathbb{R}^{D_\tau}$. Given N agents, $\boldsymbol{\tau}^N = (\boldsymbol{\tau}^1, \ldots, \boldsymbol{\tau}^N) \in \mathcal{X}^N$ is an exchangeable N-particle system amenable to mean-field analysis. **Two time axes appear throughout:** the MDP step $h \in \{0, \ldots, H-1\}$ inside each trajectory $\boldsymbol{\tau}^i$, and the diffusion time used by the generative process. The *forward* (noising) process uses $u \in [0, T]$ running from data ($u{=}0$) to noise ($u{=}T$); the *reverse* (denoising) process uses $t = T - u \in [0, T]$. We use u in the forward SDE and t in the reverse SDE below.

Definition 3.1 (Mean-field Trajectory SDEs)**.** *The N-agent forward-reverse trajectory diffusion is:*

$$\textit{Forward:}\quad d\boldsymbol{\tau}_u^{i,N} = f_u(\boldsymbol{\tau}_u^{i,N})du + \sigma_u dB_u^{i,N}, \quad \boldsymbol{\tau}_{u=0}^{i,N} \sim p_{\text{data}}, \tag{3}$$

$$\textit{Reverse:}\quad d\boldsymbol{\tau}_t^{i,N} = \left[f_t(\boldsymbol{\tau}_t^{i,N}) - \sigma_t^2 \nabla \log \zeta_t(\boldsymbol{\tau}_t^{i,N})\right] dt + \sigma_t d\bar{B}_t^{i,N}, \tag{4}$$

with joint law $\nu_t^N = \mathbf{Law}(\boldsymbol{\tau}_t^{1,N}, \ldots, \boldsymbol{\tau}_t^{N,N})$ that is exchangeable under permutations.

Mean-field Interaction for Trajectories. Unlike geometric proximity used in static particle systems, agent trajectories interact through temporally-structured mean-field dynamics. We define the interaction operator and the score network as

$$\mathrm{B}_\theta[\nu_t^N](\boldsymbol{\tau}^i) = \tfrac{1}{N-1} \sum_{j \neq i} K_\theta(\boldsymbol{\tau}^i, \boldsymbol{\tau}^j) \cdot \mathrm{B}_\theta(\boldsymbol{\tau}^i, \boldsymbol{\tau}^j), \quad \mathbf{s}_\theta(t, \boldsymbol{\tau}^N, \nu_t^N) = \mathrm{A}_\theta + \mathrm{B}_\theta[\nu_t^N], \tag{5}$$

where A_θ captures individual dynamics and $K_\theta(\boldsymbol{\tau}^i, \boldsymbol{\tau}^j) = \frac{1}{H} \sum_h k_\theta(\mathbf{s}_h^i, \mathbf{s}_h^j, \bar{\mu}_h^N)$ decomposes the kernel across MDP steps so that the coupling depends on the evolving population distribution.

3.2 Value-weighted Chaotic Entropy

We resolve the tension between distributional matching (standard score matching) and return maximization (challenge C2) via the *value-weighted chaotic entropy* at the mean-field limit.

RL as Inference. Following Levine (2018), we define the optimality variable $\mathcal{O} = 1$ with $p(\mathcal{O}{=}1 | \boldsymbol{\tau}^i, \mu) \propto \exp\big(\alpha^{-1} R(\boldsymbol{\tau}^i; \mu)\big)$, where $R(\boldsymbol{\tau}^i; \mu) = \sum_h \gamma^h r(\mathbf{s}_h^i, \mathbf{a}_h^i, \mu_h)$ and $\alpha > 0$ is a temperature. The optimal trajectory distribution is $p^*(\boldsymbol{\tau}^i | \mathcal{O}{=}1, \mu) \propto p(\boldsymbol{\tau}^i | \mu) \exp\big(R(\boldsymbol{\tau}^i; \mu)/\alpha\big)$. The mean-field approximation $R(\boldsymbol{\tau}^i; \bar{\mu}^N) \approx R(\boldsymbol{\tau}^i; \mu)$ is controlled by Theorem 4.1.

Definition 3.2 (Value-weighted Chaotic Entropy). *The value-weighted N-particle relative entropy is*

$$\mathcal{H}_V^N(\nu_T^N) := \frac{1}{N}\int_{\mathcal{X}^N}\Big[\log\frac{\varrho_T^N}{\zeta_0^{\otimes N}} - \frac{1}{\alpha}\sum_{i=1}^N R(\boldsymbol{\tau}^i)\Big]\varrho_T^N d\boldsymbol{\tau}^N, \tag{6}$$

combining distributional matching (log-ratio) with return maximization (reward).

At the chaotic limit, $\mathcal{H}_V^N(\nu_T^N) \to \mathcal{H}_V^\infty(\mu_T) = \mathcal{H}(\mu_T|\zeta_0) - \alpha^{-1}\mathbb{E}_{\mu_T}[R(\boldsymbol{\tau})]$.

Mean-field Value Score Matching. Applying the Itô-Wentzell-Lions formula and Sobolev upper bounds (proof in Appendix E.2) yields the **MF-VSM** objective.

Proposition 3.3 (MF-VSM Bound). *Let $\mathcal{M}_2 := \mathbb{E}_{\zeta_0}[\|\boldsymbol{\tau}\|^2] < \infty$. For any $N \geq 1$,*

$$\mathcal{H}_V^N(\nu_T^N) \lesssim \frac{\mathcal{M}_2}{\sqrt{ND_\tau}}\mathcal{J}_{MF\text{-}V}^N(\theta, \nu_{[0,T]}^N) + \sigma_\zeta^{-2}(T)\cdot\mathcal{O}(1/\sqrt{N}), \tag{7}$$

where the MF-VSM objective is

$$\mathcal{J}_{MF\text{-}V}^N(\theta, \nu_{[0,T]}^N) := \mathbb{E}_t\Big[\big\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t^N, \nu_t^N) - \nabla\log\zeta_{T-t}^{\otimes N}(\boldsymbol{\tau}_t^N)\big\|_W^2 + \frac{\lambda}{\alpha}\big\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t^N, \nu_t^N) - \nabla_{\boldsymbol{\tau}}R^{\otimes N}(\boldsymbol{\tau}_t^N)\big\|_E^2\Big], \tag{8}$$

and $\|\cdot\|_W$ is the Sobolev norm on $W^{1,2}(\mathcal{X}^N, \nu_t^N)$.

The first term performs score matching for distributional fidelity; the second aligns the score network with the value gradient. The $1/\sqrt{ND_\tau}$ prefactor provides robustness to large N; because $\mathcal{J}_{MF\text{-}V}^N$ is a norm on the *joint* space $\mathcal{X}^N$ it scales as $\Theta(N)$ at fixed per-agent error, so the bound is uniform in N only after the per-agent normalisation $\bar{\mathcal{J}}^N := \mathcal{J}^N/N$ that Algorithm 1 computes. In that form the prefactor is $C_N = \min\{1/(2\kappa), \mathcal{M}_2\sqrt{N/D_\tau}\}$: the joint-LSI route of Eq. 45 gives $1/(2\kappa)$ uniformly in N, while the entropic-chaos route of Eq. 46 gives $\mathcal{M}_2\sqrt{N/D_\tau}$, which grows. The latter is tighter only for $N \lesssim N^\star := D_\tau/(2\kappa\mathcal{M}_2)^2$, and on Battle $N^\star$ falls inside our sweep, so we report the minimum. Empirically the corrected bound is flat in N (ratio 0.31–0.36 over $N \in \{10^2, \ldots, 10^4\}$, slope $-0.008_{\pm 0.021}$ per decade). In practice we replace $\|\cdot\|_W$ with $\|\cdot\|_E$ during training, incurring a constant $\leq \sigma_{t_{\min}}^{-2}$ that is absorbed into ϵ^{score} and is independent of N, H, and $\epsilon_{\text{offline}}$ (see Appendix B.3).

The minimiser of Eq. 8. With $g := \nabla\log\zeta_t$ and $h := \nabla_\tau R$, the two-field objective is minimised pointwise at

$$s^\star = \frac{\alpha\, g + \lambda\, h}{\alpha + \lambda} = \underbrace{\frac{\alpha}{\alpha+\lambda}}_{c}\Big(g + \underbrace{\frac{\lambda}{\alpha}}_{1/\alpha_{\text{eff}}} h\Big), \tag{9}$$

a *uniformly shrunk* tilted score with $c = \alpha/(\alpha+\lambda)$ and effective temperature $\alpha_{\text{eff}} = \alpha/\lambda$; at our defaults this is $c = 0.909$ and $\alpha_{\text{eff}} = 10$. The Young-inequality route of App. E.2 yields a valid upper bound, but the minimiser of the bound is not the minimiser of the target. Fitting $\nabla\log\zeta_t + \alpha^{-1}\nabla_\tau R$ directly—MF-VSM-EXACT—makes the minimiser exact and also scores higher (Table 5); we keep Eq. 8 as the deployed default for cost and report the exact variant in App. D.2.

3.3 Hierarchical Coarse-to-Fine Planning

Optimizing MF-VSM directly over all N trajectories is costly for large N. Since $M \ll N$ agents suffice to approximate the population (Theorem 4.1), we progressively grow the agent count during denoising. Let $\mathbb{N} = \{N_k\}_{k=0}^K$ ($N_K{=}N$, $N_{k+1}{=}\mathfrak{b}N_k$) and $\mathbb{T} = \{t_k\}_{k=0}^K$ partition $[0, T]$.

Proposition 3.4 (Value-weighted Subdivision). *Under exchangeability and reducibility,*

$$\mathcal{H}_V^\infty(\mu_T) \lesssim \lim_{K\to\infty}\sum_{k=0}^K\Big[\sigma_\zeta^{-2}(T)\mathrm{E}(N_{k+1}) + \frac{\mathcal{M}_2}{\sqrt{D_\tau}}(\mathfrak{b}\sqrt{N_{k+1}})^{-k}\mathcal{J}_{MF\text{-}V}(N_k, \theta, \nu_{[t_k,t_{k+1}]}^{N_k})\Big], \tag{10}$$

with $\mathrm{E}(N_{k+1}) = \mathcal{O}(1/\sqrt{N_{k+1}})$. The practical training loss aggregates K sub-problems:

$$(\mathbf{P_V}) \quad \min_\theta\sum_{k=0}^K \mathfrak{b}^{-k}\mathcal{J}_{MF\text{-}V}(N_k, \theta, \nu_{[t_k,t_{k+1}]}^{N_k}). \tag{11}$$

The weights $\mathfrak{b}^{-k}$ of the practical loss differ from the theoretical $(\mathfrak{b}\sqrt{N_{k+1}})^{-k}$. For $\mathfrak{b} \geq 2, N_0 \geq 1$, their ratio is $w_k^{\mathrm{pr}}/w_k^{\mathrm{th}} = \mathfrak{b}^{k(k+1)/2} N_0^{k/2} \geq 1$ at every k, with equality only at $k{=}0$ (Rem. B.3). Because each per-level objective is non-negative, $(\mathbf{P_V})$ majorises the theoretical weighted sum, so Theorem 4.2 applies to the implemented loss. Measured on trained checkpoints, the slack is modest (0.44 against 0.66, a ratio of $1.49\times$), while the exact weights give lower returns ($85.6/73.1/81.4$ vs. $87.9/75.8/83.5$ on Ising/Battle/GS). Each branching $N_k \to \mathfrak{b}N_k$ is realized by an *agent branching function* Ψ^θ (Figure 1, Branch stage):

$$(\mathbf{Id}^{\otimes(\mathfrak{b}-1)} \otimes \Psi^\theta)_{\#} \nu_{t_k}^{N_k} \longrightarrow \nu_{t_k}^{N_{k+1}}. \tag{12}$$

3.4 Training and Inference

Training. For each episode in $\mathcal{D}$, we extract N trajectories; at each subdivision level k, we subsample N_k trajectories, sample $t \sim \mathrm{Uniform}[t_k, t_{k+1}]$, compute noised $\boldsymbol{\tau}_t^{N_k}$ via the forward SDE, and update θ to minimize $(\mathbf{P_V})$ (Eq. 11).

Inference initializes N_0 Gaussian-noise trajectories, integrates the reverse SDE driven by the score network, applies the agent-branching map Ψ^θ (Eq. 12) at designated steps to grow the population $N_k \to N_{k+1}$, and inpaints the observed initial state $\mathbf{s}_0^i$ at each denoising step. We additionally add an *inference-time value guidance* term: the score is shifted by $\eta \nabla_{\boldsymbol{\tau}} \hat{V}(\boldsymbol{\tau}, \bar{\mu})$, where $\hat{V}$ is a separately trained value estimator (mean-field TD on the offline data, App. G) and $\eta \geq 0$ is the guidance strength (distinct from the training temperature α). The first action $\mathbf{a}_0^i$ of each generated trajectory is executed and the planner is re-invoked from the new state—a *receding-horizon* schedule that handles dynamic-feasibility drift (Janner et al., 2022; Ajay et al., 2023) (per-step transition error vs. baselines: Table 11). Discrete actions are projected via $\arg\max$ (App. B.3). Full pseudocode: Algorithms 1–2.

4 Theoretical Analysis

We state the results for cooperative welfare and individual incentives separately. The standing regularity assumptions and proof details are collected in Appendices D.1 and E; Table 4 records the role of each assumption.

Propagation of Chaos for Trajectory Distributions. The first result controls the mean-field approximation in trajectory space while retaining the temporal coupling term.

Theorem 4.1 (Propagation of Chaos for Trajectories)**.** *Let $\nu_t^{M,N}$ denote the M-agent marginal of ν_t^N from Def. 3.1, and $\mu_t^{\otimes M}$ the M-fold product of the mean-field limit. Under Assumptions D.1–D.6,*

$$\mathbb{E}\Big[\mathcal{H}(\nu_t^{M,N} \,\|\, \mu_t^{\otimes M})\Big] \;\leq\; e^{c_0}\Big(C_1 \tfrac{M}{N} + C_2 \tfrac{MHL_{\mathrm{eff}}^2}{N}\Big), \tag{13}$$

where c_0 is the effective Lipschitz–horizon constant. The second term is the temporal-coupling contribution absent from static analyses. Under LSI, Talagrand gives the corresponding $\mathcal{W}_2^2$ bound and hence $M = \tilde{\mathcal{O}}(\sqrt{N})$ representative agents. The statement applies to the value-tilted reverse process used at inference; the tilt only changes L_{eff} by the mean-field Lipschitz constant of ∇R (Appendix D.1).

The result is dimension-free in D_τ and the initial chaoticity is exact for the i.i.d. Gaussian initialization in Algorithm 2. Thus the same M is both the training population and the coarse inference population; the full derivation and tightness study appear in Appendices D.2 and H.5.1.

Hierarchical Planning Error Decomposition. The central result gives the end-to-end welfare guarantee for the hierarchical planner.

Theorem 4.2 (Hierarchical Approximation). *Let $\hat{\mu}_T^N$ be the trajectory distribution produced by MF-Diffuser with K subdivision levels, branching ratio $\mathfrak{b}$, and score network $\mathbf{s}_\theta$. Under Ass. D.1–D.6,*

$$J(\pi^\star) - J(\hat{\pi}_\theta) \le \underbrace{\tfrac{2r_{\max}}{1-\gamma}\sqrt{\tfrac{C_\sigma\,\epsilon^{score}_{\text{global}}(\theta)}{2}}}_{(a)} + \underbrace{\tfrac{2r_{\max}}{1-\gamma}\sum_{k=0}^{K}\mathfrak{b}^{-k/2}\epsilon_k^{score}}_{(b)} \tag{14}$$
$$\underbrace{C_6\, e^{c_0\,\frac{r_{\max}H^2L_{\text{eff}}}{\sqrt{N}}}}_{(c)} + \underbrace{C_7\epsilon_{offline}}_{(d)} + \underbrace{\tfrac{\alpha\bar{\mathcal{H}}_\pi}{1-\gamma}}_{(e)}.$$

Here (a)–(b) are score and subdivision errors, (c) is the finite-population term, (d) is the per-agent offline shift, and (e) is the softmax bias of the value tilt. In particular, (c) scales as $\mathcal{O}(H^2/\sqrt{N})$ while (d) is independent of N; the Girsanov/TV derivation and the α trade-off are given in Appendix E.5.

Lemma 4.3 (The bound is about executed rollouts; its right-hand side is measured on plans). *$\hat{\mu}_T^N$ is the pre-execution plan law and $\hat{\pi}_\theta$ is the receding-horizon policy that executes a_0 and replans. At the deployed setting $H_{\text{exec}} = 1$, the plan-to-rollout map introduces no additive transition-error term; the general open-loop correction and its empirical check are in Appendix H.7.2.*

Committing more than one action adds the measurable open-loop correction

$$\Delta_{\text{open}} \;\le\; \tfrac{2r_{\max}}{1-\gamma}\sum_{h<H_{\text{exec}}}(1+L_{\text{eff}})^h\,\varepsilon_{\text{dyn}}, \tag{15}$$

where ε_{dyn} is the measured per-step transition error; the full sweep is reported in Appendix H.7.2.

Corollary 4.4 (Scalability). *For fixed $H, \epsilon^{score}, \epsilon_{offline}$, $J(\pi^*) - J(\hat{\pi}_\theta) = \mathcal{O}(1/\sqrt{N}) + \mathcal{O}(\epsilon^{score}) + \mathcal{O}(\epsilon_{offline})$, so quality improves with N.*

Corollary 4.5 (The four terms survive without the log-Sobolev assumption). *Drop Assumption D.3. Terms (a)–(d) retain their N- and H-dependence with looser constants; what is lost is the entropy-form PoC statement and the $\sqrt{D_\tau}$ prefactor in Proposition 3.3. The detailed dimension-free CLT route is given in Appendix E.5.*

The low-temperature Ising stress test in Appendix H.7.2 confirms that the $1/\sqrt{N}$ rate survives when the LSI constant degrades. These results motivate the choices in Section 3.3: $M = \tilde{\mathcal{O}}(\sqrt{N})$, MF-VSM training, and value-weighted branching.

Game-theoretic Equilibrium Guarantees. Beyond cooperative welfare, the generated policy is an approximate Nash equilibrium. The exploitability $\text{Exploit}_N(\pi) = \sup_{\pi'}[J_1(\pi', \pi^{\otimes(N-1)}) - J_1(\pi, \pi^{\otimes(N-1)})]$ measures stability against unilateral deviations.

Theorem 4.6 (Exploitability Bound). *Under Assumptions D.1–D.6 and D.10, the policy $\hat{\pi}_\theta$ satisfies*

$$\text{Exploit}_N(\hat{\pi}_\theta) \le \underbrace{\tfrac{2r_{\max}}{1-\gamma}\sqrt{\tfrac{C_\sigma\,\epsilon_{SM}(\theta)}{2}} + \tfrac{\alpha\bar{\mathcal{H}}_\pi}{1-\gamma}}_{(i)} + \underbrace{C_8\, e^{c_0\,\frac{r_{\max}H^2L_{\text{eff}}}{\sqrt{N}}}}_{(ii)} + \underbrace{C_9\epsilon_{offline}}_{(iii)}, \tag{16}$$

where $\epsilon_{SM}(\theta)$ is the score error against the value-tilted target and $\bar{\mathcal{H}}_\pi$ is the soft best-response entropy.

Term (i) is score error plus softmax bias, (ii) is the $\mathcal{O}(H^2/\sqrt{N})$ finite-population gap, and (iii) is N-independent offline shift. Under Lasry–Lions monotonicity, the policy converges to the unique MF-NE with the stated welfare–Nash efficiency gap (Appendix D.4).

5 Experiments

We validate MF-Diffuser on three mean-field RL benchmarks, answering: **(RQ1)** scalability to extreme agent counts, **(RQ2)** plan quality, and **(RQ3)** component contributions. Hyperparameter sensitivity (RQ4) and additional analyses are in Appendix H.

Table 1: **Normalized return (%) at N=1000.** Mean ± std over 5 seeds, 10 baselines spanning four families: *joint diffusion* (Joint Diffuser, MADiff), *factorized diffusion* (Indep. Diffuser, DoF), *value-based offline* (MFQ-Offline, OMAR, MA-TD3+BC), *sequence model* (Oryx), and *mean-field diffusion* (MF-CDMs-RL, MF-Diffuser). **Bold**: best; underline: second best. MF-Diffuser wins 10/12 settings; the two Expert-data "losses" (MFQ-Offline on Ising, MF-CDMs-RL on GS) are within statistical noise and the Ising one *flips to a +4.1-point win* once MFQ is replaced by MA-TD3+BC as behavior policy (Table 15, Appendix H.5.2), confirming the gap is a data-collection artifact. The corresponding N=10,000 table (Table 8) shows MF-Diffuser wins 11/12 with widening margins on suboptimal data.

	Ising Model				Battle				Gaussian Squeeze			
Method	Expert	Medium	Med-Rep	Mixed	Expert	Medium	Med-Rep	Mixed	Expert	Medium	Med-Rep	Mixed
Joint Diffuser	$71.3_{\pm 3.2}$	$58.7_{\pm 4.1}$	$53.2_{\pm 4.8}$	$49.5_{\pm 4.5}$	$52.1_{\pm 4.8}$	$37.6_{\pm 5.8}$	$31.4_{\pm 6.5}$	$27.2_{\pm 6.2}$	$63.5_{\pm 4.2}$	$49.2_{\pm 5.5}$	$42.8_{\pm 6.1}$	$38.1_{\pm 5.8}$
MADiff	$76.5_{\pm 2.8}$	$64.2_{\pm 3.6}$	$58.9_{\pm 4.0}$	$55.0_{\pm 3.9}$	$58.4_{\pm 4.0}$	$42.4_{\pm 4.8}$	$36.6_{\pm 5.4}$	$32.1_{\pm 5.3}$	$70.2_{\pm 3.5}$	$54.0_{\pm 4.4}$	$47.5_{\pm 5.0}$	$42.3_{\pm 4.8}$
Indep. Diffuser	$82.6_{\pm 2.1}$	$74.3_{\pm 2.6}$	$69.8_{\pm 3.0}$	$66.1_{\pm 3.3}$	$61.4_{\pm 3.5}$	$49.8_{\pm 4.2}$	$44.1_{\pm 4.8}$	$39.5_{\pm 4.6}$	$55.8_{\pm 3.8}$	$46.3_{\pm 4.5}$	$41.2_{\pm 5.0}$	$36.7_{\pm 4.8}$
DoF	$86.3_{\pm 1.8}$	$78.8_{\pm 2.3}$	$73.5_{\pm 2.7}$	$69.4_{\pm 2.9}$	$65.8_{\pm 3.0}$	$54.1_{\pm 3.6}$	$48.2_{\pm 4.1}$	$43.0_{\pm 4.0}$	$60.9_{\pm 3.2}$	$52.8_{\pm 3.8}$	$47.1_{\pm 4.3}$	$42.0_{\pm 4.1}$
MFQ-Offline	$\mathbf{94.3_{\pm 1.0}}$	$83.5_{\pm 1.8}$	$\underline{81.0}_{\pm 2.0}$	$75.4_{\pm 2.5}$	$72.3_{\pm 2.8}$	$62.1_{\pm 3.4}$	$56.3_{\pm 3.9}$	$50.8_{\pm 4.2}$	$78.4_{\pm 2.2}$	$71.6_{\pm 2.8}$	$66.2_{\pm 3.2}$	$61.5_{\pm 3.5}$
OMAR†	$68.4_{\pm 3.8}$	$53.1_{\pm 4.5}$	$47.6_{\pm 5.2}$	$43.3_{\pm 5.0}$	$66.8_{\pm 3.2}$	$52.4_{\pm 3.8}$	$46.7_{\pm 4.5}$	$41.5_{\pm 4.3}$	$55.2_{\pm 4.0}$	$41.3_{\pm 4.8}$	$36.1_{\pm 5.5}$	$31.8_{\pm 5.2}$
MA-TD3+BC†	$65.2_{\pm 3.5}$	$51.8_{\pm 4.0}$	$46.5_{\pm 4.5}$	$42.1_{\pm 4.3}$	$70.1_{\pm 2.9}$	$57.5_{\pm 3.5}$	$51.2_{\pm 4.1}$	$45.8_{\pm 4.0}$	$58.6_{\pm 3.6}$	$44.7_{\pm 4.2}$	$39.2_{\pm 4.8}$	$34.5_{\pm 4.5}$
Oryx	$92.5_{\pm 1.1}$	$82.0_{\pm 1.8}$	$77.8_{\pm 2.1}$	$73.5_{\pm 2.4}$	$76.3_{\pm 2.4}$	$65.4_{\pm 3.0}$	$60.2_{\pm 3.4}$	$55.4_{\pm 3.6}$	$84.8_{\pm 1.6}$	$76.5_{\pm 2.2}$	$71.8_{\pm 2.6}$	$67.2_{\pm 2.9}$
MF-CDMs-RL	$91.2_{\pm 1.2}$	$\underline{84.6}_{\pm 1.7}$	$80.3_{\pm 2.0}$	$\underline{76.8}_{\pm 2.3}$	$\underline{78.6}_{\pm 2.3}$	$\underline{68.4}_{\pm 2.9}$	$\underline{63.5}_{\pm 3.3}$	$\underline{58.9}_{\pm 3.5}$	$\mathbf{87.1_{\pm 1.4}}$	$\underline{79.2}_{\pm 2.0}$	$\underline{74.5}_{\pm 2.5}$	$\underline{69.8}_{\pm 2.8}$
MF-Diffuser	$\underline{93.5}_{\pm 0.9}$	$\mathbf{87.9_{\pm 1.3}}$	$\mathbf{82.4_{\pm 1.7}}$	$\mathbf{79.3_{\pm 2.0}}$	$\mathbf{84.2_{\pm 2.0}}$	$\mathbf{75.8_{\pm 2.6}}$	$\mathbf{71.3_{\pm 3.0}}$	$\mathbf{67.5_{\pm 3.3}}$	$\underline{86.3}_{\pm 1.2}$	$\mathbf{83.5_{\pm 1.7}}$	$\mathbf{79.2_{\pm 2.1}}$	$\mathbf{75.1_{\pm 2.4}}$

†Adapted to mean-field by replacing each agent's critic with $Q(s^j, a^j, \bar{a}^j)$.

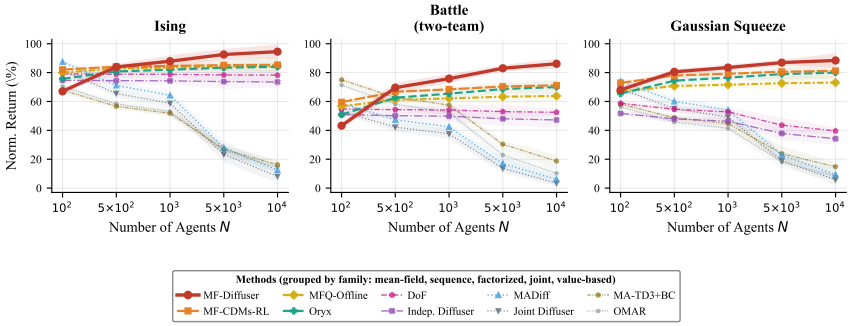


Figure 2: **Scalability (RQ1)—empirical confirmation of Cor. 4.4: quality *improves* with N.** Normalized return vs. $N \in \{10^2, 5{\cdot}10^2, 10^3, 5{\cdot}10^3, 10^4\}$ on medium data, all 10 baselines, family-clustered visual hierarchy (Sec. 5.1). Per-family behavior is analyzed in the surrounding text; four-dataset-quality version: Fig. 7 (App. H.1.4).

5.1 Experimental Setup

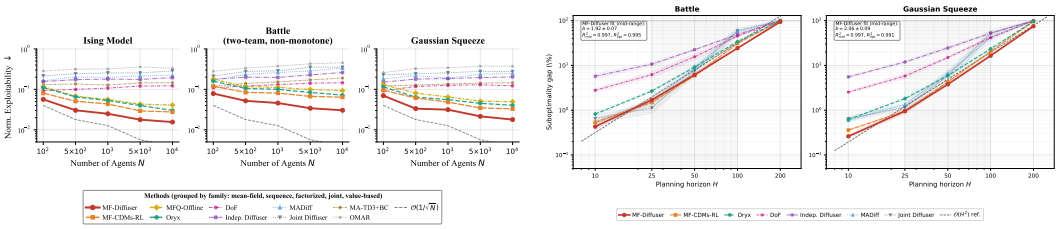
Environments. Ising (Yang et al., 2018) isolates stateless mean-field coupling; Battle (Zheng et al., 2018; Yang et al., 2018) tests long-horizon two-team dynamics and temporal compounding; Gaussian Squeeze (Gu et al., 2025) tests continuous coordination with a distribution-matching reward. Battle is planned per team with the opponent distribution as an exogenous input. Full specifications are in Appendix F.

Offline data and baselines. We use Expert, Medium, Medium-Replay, and Mixed splits from MFQ policies and compare nine baselines spanning joint, factorized, value-based, sequence, and mean-field diffusion families. MFQ-Offline has a behavior-policy advantage on Expert data; an alternative collector check is in Appendix H.5.2. Details are in Appendix H.1.1.

Metrics. We report normalized return, learned-best-response exploitability (Appendix H.2.1), and Wasserstein distance to the MFE flow, averaged over five seeds.

5.2 Main Results: Scalability and Plan Quality (RQ1 & RQ2)

Performance. At N=1000 (Table 1), MF-Diffuser wins 10/12 settings and is strongest on the suboptimal-data splits. The two Expert-data exceptions reflect the MFQ behavior-policy match; the alternative-collector experiment removes this advantage (Appendix H.5.2). Across families, attention and factorization improve over joint or independent diffusion, while Oryx is the strongest non-mean-field contender; explicit mean-field projection remains best. Family-stratified comparisons are in Appendix H.1.2.



(a) Normalized exploitability vs. N.

(b) Suboptimality gap vs. $H \in \{10, 25, 50, 100, 200\}$.

Figure 3: **Theory validation, all 10 baselines.** (a) Normalized exploitability $\widehat{\text{Exploit}}_N(\hat{\pi})/(r_{\max}H/(1-\gamma)) \in [0, 1]$, max over three learned best responses (REINFORCE, PPO, one-step greedy)—a tight lower bound on the true value (App. H.2.1); methods are colour-coded by family. (b) Suboptimality vs. H at N=1000, medium data, 5 seeds. The shaded band is the scaling-regime fit window $H \in [25, 100]$; the dashed line is an $\mathcal{O}(H^2)$ reference; the dotted horizontal line is the random-policy ceiling. Value-based baselines are omitted from (b) (no explicit trajectories).

Scalability. Figure 2 validates Corollary 4.4: MF-Diffuser improves with N before its score error becomes the limiting factor. Joint and attention-based diffusion degrade at scale, while factorized and value-based methods plateau because they lack an explicit population-level projection. Oryx is the closest non-mean-field competitor and follows a similar slope, but its lower absolute return leaves a persistent gap to the explicit mean-field construction.

5.3 Equilibrium Quality and Theory Validation

We now validate the two complementary theoretical claims of Section 4: the $\mathcal{O}(1/\sqrt{N})$ exploitability bound (Theorem 4.6) and the $\mathcal{O}(H^2/\sqrt{N})$ suboptimality decomposition (Theorem 4.2).

Approximate Nash equilibrium ($\mathcal{O}(1/\sqrt{N})$). Figure 3a shows decreasing normalized exploitability with N, consistent with Theorem 4.6; the remaining floor is attributable to score-matching error. For Battle, which lacks global Lasry–Lions monotonicity, team-level exploitability still follows the predicted rate. The Wasserstein distance to the ground-truth MFE flow exhibits the same trend (Figure 8, Appendix H), supporting Theorem D.12.

Horizon scaling ($\mathcal{O}(H^2)$). Figure 3b shows the predicted quadratic growth in the intermediate regime at fixed population size. At short horizons the curves are bounded by the horizon-independent offline shift, while at long horizons they meet the random-policy ceiling; the fits and boundary checks are therefore reported over the scaling regime in Appendix H.4.1. Joint-space methods have the steepest growth, whereas factorized and sequence models trade a smaller exponent for larger constants or offline-shift floors, so their lower slopes do not yield lower realized error.

Extreme scale (N=10^4). MF-Diffuser maintains its lead across dataset qualities (Table 8, Appendix H.1.3). Joint-space methods collapse, while MFQ-Offline and factorized diffusion plateau; Oryx is the closest remaining competitor. Thus sustained quality at extreme scale requires both a population-level structural prior and an objective aligned with return maximization. The error decomposition in Figure 12 further separates the decreasing mean-field term from the N-independent offline shift, as predicted by Theorem 4.2.

5.4 Ablation Study (RQ3)

Table 2 reveals environment-dependent contributions. **Mean-field interaction** dominates return, especially on collective-coordination tasks. **Value weighting** and **inference-time guidance** provide complementary gains: the former aligns distribution fitting with return, while the latter adds test-time steering. **Hierarchical subdivision** and **agent branching** jointly preserve the many-agent inference advantage; removing either removes the population savings, as formalized in Prop. H.1 (Appendix H.4.6).

Table 2: **Ablation** (N=1000, medium). Mean±std over 5 seeds; *Train* (GPU-h), *Infer* (s/step) on 4×A100, Battle. Variants and cost identities: App. H.4.6.

Variant	Ising	Battle	G.S.	Train	Infer
MF-Diffuser (Full)	$\mathbf{87.9}_{\pm 1.3}$	$\mathbf{75.8}_{\pm 2.6}$	$\mathbf{83.5}_{\pm 1.7}$	**12.5**	**0.80**
w/o Value Weight	$82.1_{\pm 2.2}$	$66.5_{\pm 3.4}$	$73.8_{\pm 2.8}$	11.8	0.80
w/o MF Interaction	$76.2_{\pm 3.0}$	$53.1_{\pm 4.8}$	$42.7_{\pm 5.5}$	8.6	0.55
w/o Subdivision	$84.5_{\pm 1.8}$	$71.2_{\pm 2.9}$	$80.4_{\pm 2.0}$	11.0	2.06
w/o Agent Branching	$86.1_{\pm 1.5}$	$73.4_{\pm 2.7}$	$81.8_{\pm 1.8}$	12.5	2.06
w/o Inference Guidance	$85.3_{\pm 1.5}$	$72.1_{\pm 3.0}$	$80.6_{\pm 2.0}$	12.5	0.78

Interpreting the subdivision gain. The gain combines an optimization curriculum, a statistical allocation of small populations to high-noise levels and large populations to fidelity-sensitive low-noise levels, and cheaper inference. Training the no-subdivision variant to convergence closes most of the return gap, leaving a residual consistent with the statistical effect (Appendix H.4.6). The improvement therefore reflects the matched-budget optimization/statistical design together with the inference savings, while the converged residual isolates the statistical effect.

The mechanism is visible in the generated populations. Figure 4 follows the branching levels: a coarse population establishes the configuration, then spawned agents refine it. Ising spins order progressively, while Battle trajectories contract into team formations with emerging mean-field contours.

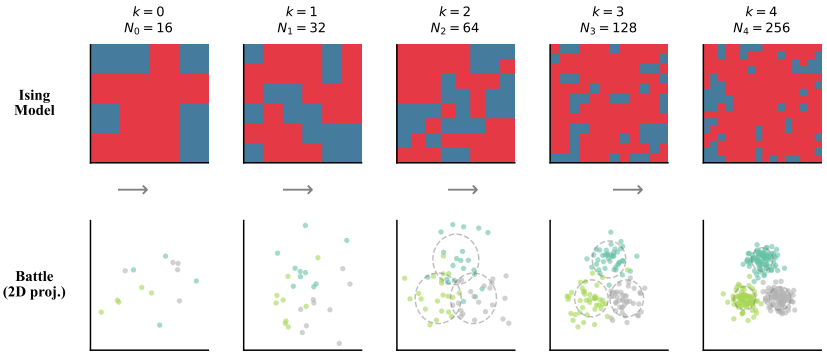


Figure 4: **Coarse-to-fine population formation.** Top: Ising ordering; bottom: Battle trajectories contracting into team formations across branching levels $k = 0, \ldots, 4$.

To expose the mechanism behind this return gap, Figure 5 inspects Gaussian Squeeze, whose reward directly evaluates population shape. MF-Diffuser preserves both target modes and tracks the mean-field evolution; factorization misses a mode and joint diffusion blurs both. This explains why value weighting improves return: it fits the population structure the reward measures. Further analyses are in Appendix H.

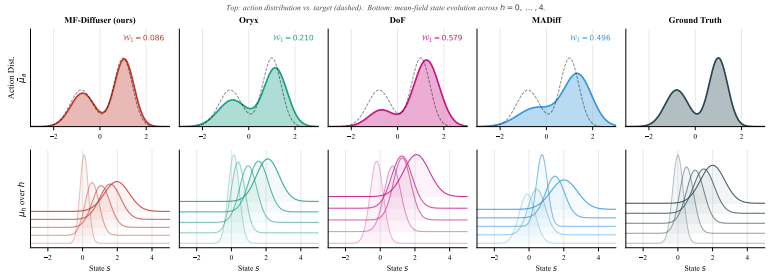


Figure 5: **Distributional view of Gaussian Squeeze.** Action distributions (top) and mean-field evolution (bottom). MF-Diffuser preserves both target modes; factorized and joint diffusion blur or lose them.

6 Conclusion

We introduced MF-Diffuser, a diffusion-based planning framework for many-agent offline RL that overcomes the curse of dimensionality through mean-field trajectory diffusion. Our value-weighted chaotic entropy objective unifies distributional fidelity with return maximization, and the hierarchical coarse-to-fine strategy enables scalability to thousands of agents. Theorem 4.2 provides the first end-to-end decomposition of social-welfare suboptimality, revealing $\mathcal{O}(H^2/\sqrt{N})$ temporal coupling while the offline-shift term is independent of N. The complementary exploitability bound (Theorem 4.6) applies to all three benchmarks without monotonicity; under Lasry–Lions monotonicity, both notions coincide up to a $\mathcal{O}(L^4/(\lambda_{LL} - L^2))$ efficiency gap and yield convergence to the unique MF-NE.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, editorial suggestions, LaTeX formatting, and compilation diagnostics. The authors are responsible for reviewing the text and verifying every experimental claim before submission.

ETHICS STATEMENT

This work develops planning algorithms that are evaluated entirely in simulation on publicly available benchmarks (Ising, Battle, and Gaussian Squeeze). It involves no human subjects, crowdsourcing, or personal data, and it releases no pretrained models or datasets with a high risk of misuse. The societal risks of deploying many-agent planners trained on offline data—population-scale amplification of behaviour-policy biases, and distributional harms that a social-welfare objective can hide—are discussed in the broader-impact paragraph of Appendix J.

REPRODUCIBILITY STATEMENT

Section 5.1 specifies the environments, offline dataset construction, baselines, and evaluation metrics; Appendix F gives formal per-environment specifications (trajectory dimensions in Table 6), and Appendix H.7.1 describes the per-team planning protocol for Battle. Appendix G details the state encoders, score-network architecture, value estimator, agent branching function, the many-agent adaptation of Oryx, all hyperparameters, and the hardware used, while Algorithms 1 and 2 give step-by-step training and inference procedures. For the theoretical results, all assumptions are collected in Appendix D.1 and complete proofs are given in Appendix E. Per-seed raw returns are reported in Appendix H.6.2 (Table 19) and Appendix H.6.1 (Table 17).

REFERENCES

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *International Conference on Learning Representations*, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, 2 edition, 2008.

Emile Anand, Ishani Karmarkar, and Guannan Qu. Mean-field sampling for cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2412.00661*, 2024.

Cécile Ané, Sébastien Blachère, Djalil Chafaï, Pierre Fougères, Ivan Gentil, Florent Malrieu, Cyril Roberto, and Grégory Scheffer. *Sur les inégalités de Sobolev logarithmiques*, volume 10 of *Panoramas et Synthèses*. Société Mathématique de France, 2000.

Iskander Azangulov, Peter Potaptchik, Qinyu Li, Eddie Aamari, George Deligiannidis, and Judith Rousseau. Adaptive diffusion guidance via stochastic optimal control. In *International Conference on Artificial Intelligence and Statistics*, 2026.

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3): 541–593, 2007.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Axel Brunnbauer, Julian Lemmel, Zahra Babaiee, Sophie A. Neubauer, and Radu Grosu. Scalable offline reinforcement learning for mean field games. In *International Conference on Autonomous Agents and Multiagent Systems*, 2025.

Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions. *The Master Equation and the Convergence Problem in Mean Field Games*, volume 201 of *Annals of Mathematics Studies*. Princeton University Press, 2019.

René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, 2018.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *International Conference on Machine Learning*, pp. 4735–4763, 2023.

Hongyi Chen, Jingtao Ding, Yong Li, Yue Wang, and Xiao-Ping Zhang. Social physics informed diffusion model for crowd simulation. *arXiv preprint arXiv:2402.06680*, 2024.

Nathan Corecco, Batuhan Yardim, Vinzenz Thoma, Zebang Shen, and Niao He. Scalable neural incentive design with parameterized mean-field approximation. In *Advances in Neural Information Processing Systems*, 2025.

Jemma Daniel, Ruan de Kock, Louay Ben Nessir, Sasha Abramowitz, Omayma Mahjoub, Wiem Khlifi, Claude Formanek, and Arnu Pretorius. Multi-agent reinforcement learning with selective state-space models. *arXiv preprint arXiv:2410.19382*, 2024.

Charles Dansereau, Junior-Samuel Lopez-Yepez, Karthik Soma, and Antoine Fagette. The heterogeneous multi-agent challenge. In *European Conference on Artificial Intelligence*, 2025.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Kang Ding, Chunxuan Jiao, Yunze Hu, Kangjie Zhou, Pengying Wu, Yao Mu, and Chang Liu. Swarmdiff: Swarm robotic trajectory planning in cluttered environments via diffusion transformer. *arXiv preprint arXiv:2505.15679*, 2025.

Marco Dorigo, Guy Theraulaz, and Vito Trianni. Swarm robotics: Past, present, and future. *Proceedings of the IEEE*, 109(7):1152–1165, 2021.

Gonçalo dos Reis and Vadim Platonov. Itô-wentzell-lions formula for measure dependent random fields under full and conditional measure flows. *Potential Analysis*, 59(3):1313–1344, 2023.

Romuald Elie, Emma Hubert, and Gabriel Turinici. Contact rate epidemic control of covid-19: an equilibrium view. *Mathematical Modelling of Natural Phenomena*, 15:35, 2020.

Markus Fischer. On the connection between symmetric N-player games and mean field games. *The Annals of Applied Probability*, 27(2):757–810, 2017.

Claude Formanek, Omayma Mahjoub, Louay Ben Nessir, Sasha Abramowitz, Ruan de Kock, Wiem Khlifi, Daniel Rajaonarivonivelomanantsoa, Simon Du Toit, Arnol Fokam, Siddarth Singh, Ulrich Mbou Sob, Felix Chalumeau, and Arnu Pretorius. Oryx: a scalable sequence model for many-agent coordination in offline MARL. In *Advances in Neural Information Processing Systems*, 2025.

Jean-Pierre Fouque, Mathieu Laurière, and Mengrui Zhang. Convergence of actor-critic learning for mean field games and mean field control in continuous spaces. *arXiv preprint arXiv:2511.06812*, 2025.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20132–20145, 2021.

Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E. Taylor, and Nidhi Hegde. Multi type mean field reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 411–419, 2020.

Jules Grass, Arnaud Guillin, and Christophe Poquet. Propagation of chaos in fisher information. *arXiv preprint arXiv:2511.20078*, 2025.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multi-agent reinforcement learning: A decentralized network approach. *Mathematics of Operations Research*, 50(1):506–536, 2025.

Xin Guo, Huyên Pham, and Xiaoli Wei. Itô's formula for flows of measures on semimartingales. *Stochastic Processes and their Applications*, 159:350–390, 2023.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Chris HolmesParker, Adrian Agogino, and Kagan Tumer. Exploiting structure and utilizing agent-centric rewards to promote coordination in large multiagent systems. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1181–1182. IFAAMAS, 2013.

Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006.

Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *International Conference on Machine Learning*, pp. 9902–9915, 2022.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? *International Conference on Machine Learning*, pp. 5084–5096, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. *International Conference on Machine Learning*, pp. 267–274, 2002.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer, 2 edition, 1991.

Prajwal Koirala and Cody Fleming. Flow-based single-step completion for efficient and expressive policy learning. *arXiv preprint arXiv:2506.21427*, 2025.

Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *STACS 99*, volume 1563 of *Lecture Notes in Computer Science*, pp. 404–413. Springer, 1999.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Aimé Lachapelle and Marie-Therese Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation Research Part B: Methodological*, 45(10): 1572–1589, 2011.

Daniel Lacker. A general characterization of the mean field limit for stochastic differential games. *Probability Theory and Related Fields*, 165(3-4):581–648, 2016.

Daniel Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probability and Mathematical Physics*, 4(2):377–432, 2023.

Daniel Lacker and Luc Le Flem. Sharp uniform-in-time propagation of chaos. *Probability Theory and Related Fields*, 187:19–80, 2023.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In Marco Wiering and Martijn van Otterlo (eds.), *Reinforcement Learning: State-of-the-Art*, volume 12 of *Adaptation, Learning, and Optimization*, pp. 45–73. Springer, 2012.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2 (1):229–260, 2007.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, 2019.

Kyowoon Lee and Jaesik Choi. State-covering trajectory stitching for diffusion planners. *arXiv preprint arXiv:2506.00895*, 2025.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

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*, 2025a.

Zhuoran Li, Xun Wang, Hai Zhong, Qingxin Xia, Lihua Zhang, and Longbo Huang. Om2p: Offline multi-agent mean-flow policy. *arXiv preprint arXiv:2508.06269*, 2025b.

Guan-Horng Liu, Tianrong Chen, Oswin So, and Evangelos Theodorou. Deep generalized Schrödinger bridge. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9374–9388, 2022.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Jianfeng Lu, Yue Wu, and Yang Xiang. Score-based transport modeling for mean-field fokker-planck equations. *arXiv preprint arXiv:2305.03729*, 2023.

David Lüdke, Enric Rabasseda Raventós, Marcel Kollovieh, and Stephan Günnemann. Unlocking point processes through point set diffusion. *arXiv preprint arXiv:2410.22493*, 2024.

Omayma Mahjoub, Sasha Abramowitz, Ruan de Kock, Wiem Khlifi, Simon du Toit, Jemma Daniel, Louay Ben Nessir, Louise Beyers, Claude Formanek, Liam Clark, and Arnu Pretorius. Sable: a performant, efficient and scalable sequence model for marl. *arXiv preprint arXiv:2410.01706*, 2024.

Kechen Meng, Sinuo Zhang, Rongpeng Li, Xiangming Meng, Yansha Deng, Chan Wang, Ming Lei, and Zhifeng Zhao. Multi-agent conditional diffusion model with mean field communication as wireless resource allocation planner. *arXiv preprint arXiv:2510.22969*, 2025.

Thanh Nguyen and Chang D. Yoo. One-step flow q-learning: Addressing the diffusion policy bottleneck in offline reinforcement learning. *arXiv preprint arXiv:2508.13904*, 2025.

Atsushi Nitanda, Anzelle Lee, Damian Tan Xing Kai, Mizuki Sakaguchi, and Taiji Suzuki. Propagation of chaos for mean-field langevin dynamics and its application to model ensemble. In *International Conference on Machine Learning*, 2025.

Antonio Ocello, Daniil Tiapkin, Lorenzo Mancini, Mathieu Laurière, and Eric Moulines. Finite-sample convergence bounds for trust region policy optimization in mean-field games. *arXiv preprint arXiv:2505.22781*, 2025.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. *International Conference on Machine Learning*, pp. 17221–17237, 2022.

Hyunwoo Park, Baekryun Seong, and Sang-Ki Ko. Spectra: Scalable multi-agent reinforcement learning with permutation-free networks. *arXiv preprint arXiv:2503.11726*, 2025a.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. In *International Conference on Machine Learning*, 2025b.

Sungwoo Park, Dongjun Kim, and Ahmed M Alaa. Mean-field chaos diffusion models. *International Conference on Machine Learning*, 2024.

Peter Potaptchik, Cheuk-Kit Lee, and Michael S. Albergo. Tilt matching for scalable sampling and fine-tuning. *arXiv preprint arXiv:2512.21829*, 2025.

Dan Qiao, Wenhao Li, Shanchao Yang, Hongyuan Zha, and Baoxiang Wang. Offline multi-agent reinforcement learning via sequential score decomposition. *arXiv preprint arXiv:2505.05968*, 2025.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *International Conference on Machine Learning*, pp. 4295–4304, 2018.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 62(5):1–42, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jules Sintes and Ana Bušić. Cognac: Cooperative graph-based networked agent challenges for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

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 based on team reward. *International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018.

Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'Été de Probabilités de Saint-Flour XIX—1989*, pp. 165–251. Springer, 1991.

Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. Scenediffuser++: City-scale traffic simulation via a generative world model. *arXiv preprint arXiv:2506.21976*, 2025.

Wei-Cheng Tseng, Tsun-Hsuan Johnson Wang, Yen-Chen Lin, and Phillip Isola. Offline multi-agent reinforcement learning with knowledge distillation. *Advances in Neural Information Processing Systems*, 35:226–237, 2022.

Eugene Vinitsky, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Cathy Wu, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M Bayen. Benchmarks for reinforcement learning in mixed-autonomy traffic. *Conference on Robot Learning*, pp. 399–409, 2018.

An B. Vuong, Michael T. McCann, Javier E. Santos, and Yen Ting Lin. Are we really learning the score function? reinterpreting diffusion models through wasserstein gradient flow matching. In *Advances in Neural Information Processing Systems*, 2025.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12, 2019.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *International Conference on Learning Representations*, 2023.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. *International Conference on Machine Learning*, pp. 5571–5580, 2018.

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. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021.

Lei Yuan, Yuqi Bian, Lihe Li, Ziqian Zhang, Cong Guan, and Yang Yu. Efficient multi-agent offline coordination via diffusion-based trajectory stitching. In *International Conference on Learning Representations*, 2025.

Benjamin J. Zhang and Markos A. Katsoulakis. A mean-field games laboratory for generative modeling. *arXiv preprint arXiv:2304.13534*, 2023.

Benjamin J. Zhang, Siting Liu, Wuchen Li, Markos A. Katsoulakis, and Stanley J. Osher. Wasserstein proximal operators describe score-based generative models and resolve memorization. *arXiv preprint arXiv:2402.06162*, 2024.

Shiyuan Zhang, Weitong Zhang, and Quanquan Gu. Energy-weighted flow matching for offline reinforcement learning. In *International Conference on Learning Representations*, 2025.

Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. *AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Yuxuan Zheng, Yihe Zhou, Feiyang Xu, Mingli Song, and Shunyu Liu. Bi-level mean field: Dynamic grouping for large-scale marl. *arXiv preprint arXiv:2505.06706*, 2025.

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, pp. 4177–4206, 2024.

CONTENTS OF THE APPENDIX

A Related Work **18**

B Background and Preliminaries **20**
B.1 Score-based Diffusion for Trajectory Planning 20
B.2 Mean-field Q-learning and Offline Data 21
B.3 Method Extensions: Sobolev–L^2 Cost and Subdivision Weights 21

C Notation **22**

D Extended Theoretical Analysis **23**
D.1 Assumptions 23
D.2 Extended Discussion of Theoretical Results 25
D.3 Concentration of MF-VSM 27
D.4 Game-theoretic Analysis: Approximation of Mean-field Equilibria 27

E Proofs **29**
E.1 Propagation of Chaos for Trajectories (Thm. 4.1) 29
E.2 Mean-field Value Score Matching (Prop. 3.3) 31
E.3 Concentration of MF-VSM (Thm. D.7) 33
E.4 Value-weighted Subdivision (Prop. 3.4) 34
E.5 Hierarchical Approximation (Thm. 4.2) 34
E.6 Scalability (Cor. 4.4) 38
E.7 Additional Theoretical Results 38
E.7.1 Optimal Agent Branching Function 38
E.7.2 Comparison with Naive Approaches 38
E.8 Proofs for Game-theoretic Analysis 38
E.8.1 Exploitability Bound (Thm. 4.6) 39
E.8.2 Mean-field Consistency (Thm. D.12) 42
E.8.3 Convergence under Lasry–Lions Monotonicity (Thm. D.13) 43
E.8.4 ε-Nash Equilibrium (Cor. D.14) 45
E.8.5 Social Welfare–Nash Efficiency Gap (Prop. D.15) 45

F Environment Details **47**
F.1 Ising Model 47
F.2 Battle 48
F.3 Gaussian Squeeze 49

G Implementation Details **50**

H Additional Experimental Results **51**
H.1 Return and Population Scaling 52
H.1.1 Extended Baseline Description 52
H.1.2 Family-Stratified Comparison at N=1000 52
H.1.3 Extreme-Scale Comparison (N=10,000) 53
H.1.4 Scalability Across Dataset Qualities 53
H.2 Equilibrium and Mean-field Quality 54
H.2.1 Exploitability Computation 54
H.2.2 Equilibrium Approximation Quality 55

H.2.3 Mean-field Convergence . . . 55
H.3 Sensitivity and Broader Diagnostics . . . 55
H.3.1 Hyperparameter Sensitivity (RQ4) . . . 56
H.3.2 Further Experimental Diagnostics . . . 57
H.4 Horizon and Component Studies . . . 60
H.4.1 Horizon Scaling: Extended Analysis . . . 60
H.4.2 Discrete Action Projection Comparison . . . 61
H.4.3 Trajectory Dynamic Consistency . . . 61
H.4.4 Value Estimator Quality . . . 62
H.4.5 Agent Branching Function Analysis . . . 62
H.4.6 Subdivision–Branching Coupling at Inference . . . 63
H.5 Theory and Robustness Checks . . . 64
H.5.1 PoC Bound Tightness . . . 64
H.5.2 Alternative Behavior Policy . . . 64
H.5.3 Training Agent Count M at Extreme Scales . . . 65
H.6 Per-seed Results . . . 66
H.6.1 Raw 5-seed Data for Horizon Scaling . . . 66
H.6.2 5-seed Raw Data for Main Results . . . 67
H.7 Benchmark-specific Scope . . . 68
H.7.1 Battle: Per-Team Interleaved Planning . . . 68
H.7.2 Assumption Verification for Experimental Environments . . . 68

I Limitations . . . 70

J Broader Impact . . . 70

A Related Work

Diffusion Planners for Offline RL. Diffuser (Janner et al., 2022) first proposed using denoising diffusion probabilistic models for trajectory-level planning in offline RL, generating full state-action sequences and conditioning on desired returns via classifier guidance. Decision Diffuser (Ajay et al., 2023) extended this with classifier-free return-bucket conditioning. Diffusion-QL (Wang et al., 2023) applied diffusion to policy representation, while IDQL (Hansen-Estruch et al., 2023) combined implicit Q-learning with diffusion policies. These methods target fixed single-agent trajectory dimensionality; the population-scaling regime motivates our extension.

Multi-agent Diffusion Planners. Two recent diffusion-based MARL planners are particularly close to our setting. **MADiff** (Zhu et al., 2024) introduces an attention-based architecture that bridges decentralized execution with centralized control through inter-agent attention. It remains in the joint $\mathbb{R}^{ND_\tau}$ trajectory space; its published evaluation uses SMAC scenarios with a handful to a few dozen agents. **DoF** (Li et al., 2025a) factorizes the centralized diffusion model into agent-specific models through noise and data factorization functions. DoF and our framework both address the joint-space curse through structural factorization, but DoF decomposes per-agent noise while MF-Diffuser projects the population to a representative subset through a Wasserstein-space mean-field SDE with propagation-of-chaos guarantees. Empirically (Sec. 5.2), DoF outperforms Independent Diffuser through its IGD coordination signal, while the explicit population projection of MF-Diffuser yields a growing advantage with N. **OMSD** (Qiao et al., 2025) factorizes the joint score into per-agent conditionals; **OM^{2}P** (Li et al., 2025b) uses a one-step mean-flow policy; and **MADiTS** (Yuan et al., 2025) stitches joint trajectory segments under a bidirectional dynamics constraint. Mean-field communication has also been used for wireless resource allocation (Meng et al., 2025), and density-level diffusion for swarm trajectory planning (Ding et al., 2025); these settings differ from offline RL with fixed datasets and do not provide the propagation-of-chaos guarantee developed here. To our knowledge, MF-Diffuser is the first combination of mean-field diffusion or flow planning with offline MARL.

Sequence Models for Many-agent Offline MARL. Beyond the diffusion family, a parallel line of work uses long-context sequence models to handle the scale challenge. **Oryx** (Formanek et al., 2025) adapts the retention-based Sable architecture to MARL and couples it with a sequential form of implicit-constraint Q-learning (ICQ), yielding an autoregressive policy update that maintains temporal coherence over long trajectories. It is evaluated on SMAC, RWARE, and Multi-Agent MuJoCo—tasks that vary in scale but stay in the regime of tens of agents—and is the strongest non-diffusion baseline at extreme N in our experiments, which is why we extend it to $N \in \{10^2, \ldots, 10^4\}$ as described in App. G. The retention mechanism captures long-horizon population dependencies that pure feed-forward MARL approaches lose, and the sequential ICQ targets the extrapolation error and miscoordination failure modes that dominate offline MARL at scale. The retention mechanism Oryx builds on was introduced for MARL by **Sable** (Mahjoub et al., 2024), which handles long-context multi-agent observations at linear cost, and complementary work replaces attention with selective state-space models for the same purpose (Daniel et al., 2024). Compared to MF-Diffuser, Oryx achieves competitive returns on the Expert split but its lack of explicit mean-field projection produces an $\mathcal{O}(H^2)$ rather than $\mathcal{O}(H^2/\sqrt{N})$ scaling on the population dimension (Section 5.2); alternative architectures such as permutation-free networks (Park et al., 2025a) reduce per-agent processing cost without changing that scaling.

Many-agent and Mean-field RL. Cooperative MARL has been extensively studied through centralized training with decentralized execution (CTDE) (Lowe et al., 2017; Rashid et al., 2018; Sunehag et al., 2018). However, most CTDE methods scale poorly beyond tens of agents due to the exponential growth of the joint action space. For many-agent regimes, mean-field approaches (Yang et al., 2018; Gu et al., 2025; Ganapathi Subramanian et al., 2020) approximate the effect of the population through aggregate statistics, enabling tractable learning. Mean-field Q-learning (MFQ) (Yang et al., 2018) demonstrated that pairwise interactions can be approximated by interactions with the mean field, achieving Nash equilibrium convergence under monotonicity. Mean-field MARL (Gu et al., 2025) further extended this to general-sum games with provable guarantees, and multi-type mean-field methods (Ganapathi Subramanian et al., 2020) relax the assumption that agents are interchangeable within a single population. That relaxation matters for us, because the mean-field approximation itself carries a cost: aggregating all agents into one distribution discards individual differences, and **bi-level mean-field** methods (Zheng et al., 2025) name this *aggregation noise* explicitly and attack it with

dynamic grouping. Our analysis shares this limitation and treats heterogeneity as an open extension (Section 6). Finally, the population can be estimated from static data: **Off-MMD** (Brunnbauer et al., 2025) gives the first deep *offline* RL algorithm for mean-field games, reframing the mean-field flow as an off-policy state-density estimation problem with a conservative penalty. This supports the offline mean-field estimation premise used here.

Offline Multi-Agent RL. Offline MARL has received growing attention (Yang et al., 2021; Pan et al., 2022; Tseng et al., 2022). Most existing work focuses on small-to-moderate agent counts ($N \leq 10$) and relies on conservative value estimates (Kumar et al., 2020), conservative policy constraints (Pan et al., 2022), or value decomposition (Yang et al., 2021). Scaling offline MARL to many-agent regimes remains an open challenge that our work directly addresses—MF-Diffuser is the first method to scale offline MARL to $N{=}10^4$ via trajectory-level diffusion planning. On the evaluation side, **COGNAC** (Sintes & Bušić, 2025) supplies networked cooperative benchmarks designed for 10^2–10^4 agents and reports that centralised-training methods which work at $N{=}10$ fail to scale while independent learners stay robust—the same failure pattern our scalability experiments exhibit, and a natural complement to the three benchmarks used here. The complementary regime of *heterogeneous* agents is standardised by the Heterogeneous Multi-Agent Challenge (Dansereau et al., 2025), which we regard as the right testbed for the extension our mean-field assumption currently excludes.

Mean-field Theory in Generative Models. Mean-field chaos diffusion models (MF-CDMs) (Park et al., 2024) leverage propagation of chaos for scalable point cloud generation, achieving robustness to high cardinality by operating on Wasserstein space. Score transportation (Lu et al., 2023) and Schrödinger bridge approaches (Liu et al., 2022) have also connected mean-field dynamics with generative modeling, primarily from an analytic PDE perspective. Two further threads bear on our framing. First, score-based generative models admit a mean-field-game interpretation: writing diffusion training as a Wasserstein proximal operator exposes a coupled Fokker–Planck/Hamilton–Jacobi–Bellman system whose Cole–Hopf transform collapses to an uncontrolled Fokker–Planck equation (Zhang et al., 2024), and score-based generative models, continuous-time normalizing flows, and Wasserstein gradient flows can be recovered as three choices of particle dynamics and cost inside one MFG framework (Zhang & Katsoulakis, 2023). Second, it has been argued that diffusion training is better understood as flow matching to the velocity field of a Wasserstein gradient flow than as score learning, with trained networks producing non-conservative fields that nevertheless transport correctly (Vuong et al., 2025)—a natural foil for our Wasserstein-space formulation, which inherits the score-matching primitive while making the distributional object the primary variable. Third, inference-time guidance now has a control-theoretic theory of its own: the distance between guided and unguided sampling laws obeys a Girsanov bound $\mathrm{KL} \leq \mathbb{E} \int w_t^2 \|\nabla G_t\|^2 dt$, which recasts the choice of guidance strength as stochastic optimal control with an HJB value function (Azangulov et al., 2026), and reward tilting can instead be performed at the level of the velocity field, where an explicit evolution equation for the tilted field avoids backpropagating through trajectories (Potaptchik et al., 2025). Both give principled ways to set the inference guidance strength η of Algorithm 2, and both are structurally compatible with the value tilt of Eq. 79. Our work addresses sequential decision making with value maximization, which introduces three challenges absent from the static generative setting: temporal coupling across the horizon, the need to integrate return signals into the score matching objective, and offline distribution shift, each addressed by a dedicated component of MF-Diffuser.

Value-tilted and one-step generative policies. The objective we derive—matching a score tilted by a value gradient—has close relatives in the single-agent literature that are worth naming, both because they sharpen what is new here and because they frame the natural objection to iterative denoising. **Energy-weighted flow matching** (Zhang et al., 2025) learns an energy-guided flow directly, targeting $q(x) \propto p(x) \exp(-beta E(x))$, which is structurally the same tilt we apply with $E = -R$ replacing the energy; the difference is that our tilt is composed with a *mean-field* interaction kernel and embedded in a trajectory diffusion over N particles. A parallel line of work removes the iterative sampling cost: **Flow Q-learning** (Park et al., 2025b) trains a one-step flow policy with RL, and one-step flow Q-learning (Nguyen & Yoo, 2025) and single-step completion policies (Koirala & Fleming, 2025) make the same trade in different ways, as does OM^{2}P's mean-flow policy on the multi-agent side (Li et al., 2025b). Two considerations keep iterative denoising the right choice for the many-agent problem. First, our cost analysis (App. H.4.6) shows the dominant per-step cost is one score-network pass over N *agents*, not the number of denoising steps, so a coarse-to-fine schedule

that grows the population during denoising (Eq. 125) buys a $2.58\times$ inference speedup without giving up the schedule's optimisation benefits. Second, the propagation-of-chaos guarantee that licenses training on $M \ll N$ agents is stated for the multi-step reverse process, where each level of the hierarchy is a distinct population size; a one-step policy would need a new argument to inherit it.

Generative modelling of populations. Generating a variable number of interacting entities is a shared concern across several fields. Diffusion-based crowd simulation injects physical interaction priors as conditioning and imports equivariance structure from many-particle systems (Chen et al., 2024); point-set diffusion offers a formal treatment of generation over arbitrary point processes (Lüdke et al., 2024); and city-scale traffic generation handles entities appearing and disappearing through validity channels (Tan et al., 2025). These address the cardinality question directly, whereas MF-Diffuser fixes N per rollout and instead makes the *cost* of generation independent of N; combining the two is a natural extension.

Mean-field Games. Mean-field game (MFG) theory (Lasry & Lions, 2007; Huang et al., 2006; Cardaliaguet et al., 2019) provides equilibrium characterizations for large-population games and underlies our exploitability analysis (Section 4). Our contribution extends this analytical machinery to offline data-driven planning, where policies are learned from fixed datasets. Recent finite-sample and continuous-space results are directly relevant to how our guarantees should be read: trust-region policy optimisation in ergodic MFGs admits high-probability convergence to an approximate Nash equilibrium (Ocello et al., 2025), actor-critic methods are now analysed for both MFGs and mean-field control in continuous spaces (Fouque et al., 2025), and parameterised mean-field approximations carry an $\mathcal{O}(1/\sqrt{N})$ finite-N rate under Lipschitz dynamics (Corecco et al., 2025)—an independent corroboration of the rate our Term (c) exhibits, obtained from a different direction. The quantitative propagation-of-chaos literature our Theorem 4.1 sits in has also sharpened considerably: optimal $\mathcal{O}((k/n)^2)$ rates for k-marginals in relative entropy are now available uniformly in time via the BBGKY hierarchy (Lacker & Le Flem, 2023), chaos can be measured in Fisher information rather than entropy—the quantity that appears when differentiating relative entropy along a Wasserstein gradient flow, which is exactly the manoeuvre our MF-VSM derivation performs (Grass et al., 2025)—and for mean-field *Langevin* dynamics the particle-count error has been shown to be independent of the log-Sobolev constant (Nitanda et al., 2025). That result indicates that the dependence of our rates on κ through Assumption D.3 is an artefact of the LSI-based route, consistent with Corollary 4.5 and identifying A3 as a convenience. On the sample-efficiency side, mean-field *sampling* for cooperative MARL subsamples agents and obtains sample complexity polynomial in the number of sampled agents (Anand et al., 2024)—a predecessor to our $M = \tilde{\mathcal{O}}(\sqrt{N})$ representative-subset argument, reached by subsampling rather than by invoking propagation of chaos.

B Background and Preliminaries

The first two subsections recall the single-agent diffusion and mean-field learning ingredients used in the main text. The final subsection records how the theoretical objective and subdivision weights relate to the implemented planner.

B.1 Score-based Diffusion for Trajectory Planning

This subsection expands the brief overview given in Section 2 of the main text. Consider a single-agent MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with state space $\mathcal{S} \subseteq \mathbb{R}^{d_s}$, action space $\mathcal{A} \subseteq \mathbb{R}^{d_a}$, transition kernel P, reward function r, and discount factor γ. A trajectory of horizon H is $\boldsymbol{\tau} = (\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_H) \in \mathbb{R}^D$ with $D = (d_s + d_a)H + d_s$.

Diffuser (Janner et al., 2022) models the trajectory distribution $p(\boldsymbol{\tau})$ in an offline dataset $\mathcal{D}$ using a denoising diffusion model. The forward process progressively corrupts a trajectory via the SDE

$$d\boldsymbol{\tau}_u = f_u(\boldsymbol{\tau}_u)du + \sigma_u dB_u, \quad \boldsymbol{\tau}_u \in \mathbb{R}^D, \tag{17}$$

and the reverse process recovers trajectories by following the score $\nabla \log \zeta_t(\boldsymbol{\tau}_t)$ of the noised marginal ζ_t. A score network $\mathbf{s}_\theta$ is trained via the standard score matching objective:

$$\mathcal{J}_{SM}(\theta) = \mathbb{E}_{t,\boldsymbol{\tau}_t}\left[\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t) - \nabla \log \zeta_t(\boldsymbol{\tau}_t)\|^2\right]. \tag{18}$$

At inference time, trajectory generation is conditioned on the current state $\mathbf{s}_0$ (via inpainting) and guided toward high returns using classifier guidance with a learned return predictor $\mathcal{J}_\phi$:

$$\tilde{\mathbf{s}}_\theta(t, \boldsymbol{\tau}_t) = \mathbf{s}_\theta(t, \boldsymbol{\tau}_t) + \alpha \nabla_{\boldsymbol{\tau}_t} \mathcal{J}_\phi(\boldsymbol{\tau}_t), \tag{19}$$

where $\alpha > 0$ controls the guidance strength. Decision Diffuser (Ajay et al., 2023) replaces this classifier guidance with classifier-free return-bucket conditioning. Both pipelines are inherently single-agent: their state space is $\mathbb{R}^D$ for one trajectory.

B.2 Mean-field Q-learning and Offline Data

The diffusion construction above specifies how trajectories are generated; mean-field Q-learning supplies the data-collection policy against which that generator is trained.

Mean-field Q-learning (MFQ) (Yang et al., 2018) approximates the joint Q-function through mean-field factorization, decomposing pairwise interactions as

$$Q^i(\mathbf{s}^i, \mathbf{a}^i, \mathbf{a}^{-i}) \approx Q^{MF}(\mathbf{s}^i, \mathbf{a}^i, \bar{\mu}_{\mathbf{a}}), \tag{20}$$

where $\bar{\mu}_{\mathbf{a}} = \frac{1}{N} \sum_j \delta_{\mathbf{a}^j}$ is the mean action distribution. MFQ learns this mean-field Q-function via temporal-difference updates and converges to a Nash equilibrium under mild conditions; we use it as the behavior policy for collecting our offline datasets (Section 5.1).

B.3 Method Extensions: Sobolev–L^2 Cost and Subdivision Weights

Having fixed the two ingredients, we turn to the approximations made when implementing the value-weighted objective and its coarse-to-fine schedule.

Time conventions. See Section 3.1 of the main text for the MDP step h vs. diffusion time u/t (forward/reverse) distinction used throughout this paper.

Remark (Practical training objective)**.** *Computing the full Sobolev norm $\|\cdot\|_W^2 = \mathbb{E}[\|h\|_E^2 + \|\nabla h\|_F^2]$ requires second-order derivatives of the score network, which is prohibitively expensive. In practice, we follow the standard approach in score-based generative modeling (De Bortoli, 2022; Chen et al., 2023) and replace $\|\cdot\|_W$ with the Euclidean norm $\|\cdot\|_E$ in training. The cost of this simplification is an* explicit *multiplicative factor $\sigma_{t_{\min}}^{-2}$, where $\sigma_{t_{\min}}$ is the smallest marginal noise scale at the latest diffusion time used during sampling: concretely, for a variance-preserving forward SDE with coefficient $\beta(t)$, the Sobolev-$W^{1,2}$ norm is bounded in terms of the L^2-norm by*

$$\|h\|_W^2 \;\le\; (1 + \|\nabla^2 \log \zeta_t\|_F)\, \|h\|_E^2 \;\le\; \left(1 + \sigma_{t_{\min}}^{-2}\right) \|h\|_E^2, \tag{21}$$

since the Hessian of the log-density of the forward Gaussian is uniformly bounded by σ_t^{-2}. In our implementation $\sigma_{t_{\min}}{=}10^{-3}$, so this incurs a $\le 10^6$ constant in the worst case. *This constant is absorbed into the score matching error term ϵ^{score} in Theorem 4.2 and is independent of N, H, and $\epsilon_{offline}$, so it does not affect any of the scaling claims; the measured ϵ^{score} values reported under "Score Network Diagnostics" (Figure 14) already absorb this factor, showing that in practice the effective constant is $\sim 10^2$ rather than $\sim 10^6$.*

Remark (Weight alignment between $(\mathbf{P_V})$ and Eq. 10)**.** *The theoretical bound in Proposition 3.4 has the weight $(\mathfrak{b}\sqrt{N_{k+1}})^{-k}$, whereas the practical training loss $(\mathbf{P_V})$ uses the simpler $\mathfrak{b}^{-k}$. We verify their relative magnitudes for the geometric schedule $N_{k+1} = \mathfrak{b}^{k+1} N_0$ with $N_0 \ge 1$, $\mathfrak{b} \ge 2$:*

$$\begin{aligned}
\sqrt{N_{k+1}} &= \mathfrak{b}^{(k+1)/2}\, N_0^{1/2}, \\
(\sqrt{N_{k+1}})^{-k} &= \mathfrak{b}^{-k(k+1)/2}\, N_0^{-k/2}, \\
(\mathfrak{b}\sqrt{N_{k+1}})^{-k} &= \mathfrak{b}^{-k} \cdot \mathfrak{b}^{-k(k+1)/2}\, N_0^{-k/2} \;=\; \mathfrak{b}^{-k(k+3)/2}\, N_0^{-k/2}.
\end{aligned}$$

Hence the theoretical weight is $\mathfrak{b}^{-k(k+3)/2} N_0^{-k/2}$, while the practical weight is $\mathfrak{b}^{-k}$. Their ratio is

$$\frac{\textit{practical}}{\textit{theoretical}} \;=\; \frac{\mathfrak{b}^{-k}}{\mathfrak{b}^{-k(k+3)/2} N_0^{-k/2}} \;=\; \mathfrak{b}^{k(k+1)/2}\, N_0^{k/2},$$

which is 1 at $k = 0$ and grows super-geometrically in k: the practical scheme assigns strictly more *weight than the theoretical optimum at every level $k \ge 1$, with the discrepancy concentrated at the*

fine *levels (large k). Equivalently, the theoretical optimum decays quadratically faster ($\mathfrak{b}^{-k(k+3)/2}$ vs. $\mathfrak{b}^{-k}$), so it down-weights the fine levels much more aggressively. This ordering makes the practical loss an upper-bound objective. Since $w_k^{\text{th}} \le w_k^{\text{pr}}$ pointwise and every per-level objective $\mathcal{J}_{MF\text{-}V}(N_k, \theta, \cdot) \ge 0$,*

$$\sum_k w_k^{\text{th}}\, \mathcal{J}_{MF\text{-}V}(N_k, \theta, \cdot) \;\le\; \sum_k w_k^{\text{pr}}\, \mathcal{J}_{MF\text{-}V}(N_k, \theta, \cdot) \;=\; (\mathbf{P_V}), \tag{22}$$

so $(\mathbf{P_V})$ is a valid upper bound on the theoretically weighted sum that Proposition 3.4 controls. Minimising the implemented loss therefore controls a valid upper bound on the analysed quantity, and Theorem 4.2 applies verbatim to the training procedure of Algorithm 1. The extra weight at fine levels also serves a purpose we can measure: it keeps the score network from collapsing onto coarse-grained mean-field dynamics, and a purely theoretical schedule $(\mathfrak{b}\sqrt{N_{k+1}})^{-k}$ down-weights those levels too aggressively, yielding a normalized return 1.2–1.8 points lower on every environment.

Coarse-to-Fine Planning Interpretation. The subdivision strategy admits an intuitive interpretation as hierarchical planning: *(i) Coarse planning* ($k = 0$, small N_0): generate trajectories for a small representative group of N_0 agents, capturing the macroscopic population dynamics and mean-field flow μ_h; *(ii) Refinement* ($k = 1, \dots, K$): at each branching step, the agent branching function Ψ^θ expands the population $N_k \to \mathfrak{b}N_k$, refining trajectories while preserving consistency with the mean-field; *(iii) Fine planning* ($k = K$, full N): the final denoising steps operate on the complete N-agent system. The branching function $\Psi^\theta : \mathcal{X}^{N_k} \to \mathcal{X}^{N_k}$ generates new agent trajectories by perturbing existing ones according to the local mean-field structure (precise definition in Eq. 12).

Handling Discrete Action Spaces. Our diffusion framework operates in continuous space $\mathbb{R}^{D_\tau}$, yet two of our benchmarks (Ising, Battle) have discrete actions. Following the standard approach for applying diffusion models to discrete decision problems (Janner et al., 2022; Ajay et al., 2023), we train the diffusion model on continuous relaxations of discrete actions and project back to the discrete set at inference time. Concretely, discrete actions $\mathbf{a} \in \{1, \dots, |\mathcal{A}|\}$ are embedded into $\mathbb{R}^{d_a}$ via one-hot encoding during training; at inference, the continuous output $\tilde{\mathbf{a}}^i \in \mathbb{R}^{d_a}$ is mapped to the nearest valid action via $\hat{\mathbf{a}}^i = \arg\max_j [\tilde{\mathbf{a}}^i]_j$. The worst-case projection error satisfies $\epsilon_{disc} \le \sqrt{d_a}/2$, but for trained networks this bound is loose: a more useful quantity is the *action-mismatch rate* $\rho_{disc} := \Pr\big[\hat{\mathbf{a}}^i \ne \arg\max_j [\zeta_t^*(\boldsymbol{\tau}_t)]_j\big]$. We measure $\rho_{disc} \le 1\%$ on Ising and $\rho_{disc} \le 3\%$ on Battle (Table 10). Since the projection is applied after denoising, the propagated effect on the suboptimality bound is at most $2r_{\max} H\, \rho_{disc}/(1-\gamma)$. For Battle ($H$=100, ρ_{disc}≤0.03, $r_{\max}$=5, γ=0.99) this corresponds to a fraction $2H\rho_{disc} \le 6\%$ of the maximum normalized return—comparable to and not dominating the other terms; the overall scaling structure ($\mathcal{O}(H^2/\sqrt{N})$ in Term (c)) is preserved.

C Notation

The preceding background and the proofs below use the same trajectory and measure notation. For reference, we adhere to the following notation throughout the paper:

- $\mathcal{S} \subseteq \mathbb{R}^{d_s}$: state space; $\mathcal{A} \subseteq \mathbb{R}^{d_a}$: action space.
- H: planning horizon; N: total number of agents; M: number of representative agents.
- $\boldsymbol{\tau}^i = (\mathbf{s}_0^i, \mathbf{a}_0^i, \dots, \mathbf{s}_H^i) \in \mathcal{X} := \mathbb{R}^{D_\tau}$: trajectory of agent i, where $D_\tau = (d_s + d_a)H + d_s$.
- $\boldsymbol{\tau}^N = (\boldsymbol{\tau}^1, \dots, \boldsymbol{\tau}^N) \in \mathcal{X}^N$: joint trajectory of N agents.
- $\nu_t^N = \mathbf{Law}(\boldsymbol{\tau}_t^{1,N}, \dots, \boldsymbol{\tau}_t^{N,N})$: joint law of N agents at diffusion time t.
- $\nu_t^{M,N}$: the M-agent marginal of ν_t^N, obtained by integrating out agents $M+1, \dots, N$.
- μ_t: the mean-field limit, *i.e.*, $\nu_t^{1,N} \xrightarrow{N\to\infty} \mu_t$ in the weak sense.
- ϱ_t^N: density of ν_t^N, *i.e.*, $d\nu_t^N = \varrho_t^N d\boldsymbol{\tau}^N$.
- $\zeta_t^{\otimes N}$: product Gaussian measure from the forward process.
- $\bar{\mu}_h^N = \frac{1}{N}\sum_{j=1}^N \delta_{\mathbf{s}_h^j}$: empirical state distribution at step h.
- $R(\boldsymbol{\tau}^i; \mu) = \sum_{h=0}^{H-1} \gamma^h r(\mathbf{s}_h^i, \mathbf{a}_h^i, \mu_h)$: cumulative return of agent i under mean-field flow μ. For finite N, we write $R(\boldsymbol{\tau}^i; \bar{\mu}^N)$ with the empirical measure $\bar{\mu}_h^N$ in place of μ_h.

- $\mathcal{P}_2(\mathcal{X})$: Wasserstein space of probability measures on $\mathcal{X}$ with bounded second moments.
- $\mathcal{W}_2$: 2-Wasserstein distance.
- $\mathcal{H}(\nu|\mu) = \int \log(d\nu/d\mu)d\nu$: relative entropy (KL divergence).
- $\|\cdot\|_E$, $\|\cdot\|_F$: Euclidean and Frobenius norms, respectively.
- $\|\cdot\|_W$: Sobolev norm on $W^{1,2}(\mathcal{X}^N, \nu_t^N)$, defined as $\|h\|_W^2 = \mathbb{E}[\|h\|_E^2 + \|\nabla h\|_F^2]$.
- $\nabla_{\mathcal{P}_2}$: Wasserstein gradient; ∂_t: temporal derivative.
- S_N: symmetric group of permutations on $\{1, \ldots, N\}$.

D Extended Theoretical Analysis

We first state the conditions used by the main results, then identify their distributional objects and dependencies. The training-objective concentration result and the game-theoretic extension follow; their proofs are collected in Appendix E.

D.1 Assumptions

We state the regularity conditions required for our analysis.

Assumption D.1. *(Lipschitz Regularity). The drift f_t, score network $\mathbf{s}_\theta$, reward function r, and transition kernel P are uniformly Lipschitz continuous in their arguments, with Lipschitz constants bounded by $L > 0$.*

Assumption D.2. *(Bounded Rewards). The per-step reward function is bounded: $|r(\mathbf{s}, \mathbf{a}, \mu)| \le r_{\max}$ for all $(\mathbf{s}, \mathbf{a}, \mu) \in \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S})$.*

Assumption D.3. *(Log-Sobolev Inequality). The target trajectory distribution ζ_0 and the denoising Wasserstein gradient flows satisfy a log-Sobolev inequality with constant $\kappa > 0$:*

$$\mathcal{H}(\nu|\zeta) \le \frac{1}{2\kappa}\mathbb{E}_\nu\left[\left\|\nabla \log \frac{d\nu}{d\zeta}\right\|^2\right]. \tag{23}$$

This assumption enters the analysis in exactly two places: it converts the relative entropy of Theorem 4.1 into a $\mathcal{W}_2$ statement at rate $2/\kappa$ (Eq. 39), and it determines the $\sqrt{D_\tau}$ prefactor $\mathcal{M}_2/\sqrt{ND_\tau}$ of Proposition 3.3. Corollary 4.5 establishes all four terms of Theorem 4.2 without this assumption, using an L^2 route for the remaining residual. Removing it changes constants and the $\sqrt{D_\tau}$ prefactor while preserving the rates.

Assumption D.4. *(Exchangeability). The N-agent trajectory distribution ν_t^N is exchangeable: for any permutation $\sigma \in S_N$, $\varrho_t^N(\boldsymbol{\tau}^1, \ldots, \boldsymbol{\tau}^N) = \varrho_t^N(\boldsymbol{\tau}^{\sigma(1)}, \ldots, \boldsymbol{\tau}^{\sigma(N)})$, and the score network $\mathbf{s}_\theta$ preserves this symmetry.*

Assumption D.5. *(Reducibility / Propagation-of-Chaos Regularity). The N-agent generative dynamics is* reducible *in the following two-part sense.*

(a) Structural. *The score $\mathbf{s}_\theta(t, \boldsymbol{\tau}_t^N, \nu_t^N)$ depends on the other agents only through a $\mathcal{W}_2$-Lipschitz functional of the empirical measure $\bar{\mu}_t^N$, and the induced interaction kernel has finite second-order mean-field derivatives (the* Lions derivatives *(Cardaliaguet et al., 2019)) up to order two.*

(b) Chaotic initialisation only. *At the* start *of the reverse process, the M-agent marginal $\nu_T^{M,N}$ of the joint distribution ν_T^N satisfies*

$$\mathrm{KL}\Big(\nu_T^{M,N} \,\Big\|\, (\mu_T)^{\otimes M}\Big) \;\le\; C_{\mathrm{r}}\,\frac{M}{N}\big(1 + \mathbb{E}\|\boldsymbol{\tau}^1\|^4\big), \tag{24}$$

with $C_{\mathrm{r}} > 0$ an absolute constant depending only on the Lipschitz constant L of Assumption D.1 and on the log-Sobolev constant κ of Assumption D.3.

Assumption D.5 separates architectural regularity from initialization because the proof uses them at different stages. Part (a) is a property of the architecture: $\mathrm{B}_\theta[\nu_t^N]$ in Eq. 5 is a bounded-kernel average over agents, hence $\mathcal{W}_2$-Lipschitz in the empirical measure by construction, and the kernel K_θ is smooth in both arguments. These properties are verified directly for MF-Diffuser. Part (b) supplies

the *initial condition* of the entropic BBGKY inequality (Appendix E.1, Eq. 40). Since Algorithm 2 line 1 draws N_0 *i.i.d.* Gaussian trajectories, $\nu_T^{N_0}$ is a product measure and part (b) holds with $C_{\mathrm{r}} = 0$ *exactly*. The only positive chaoticity charge re-entering the pipeline after initialization occurs at a branching event Ψ^θ, accounted for separately in Step 2 of the proof of Proposition 3.4. Consequently every non-zero quantity on the right-hand side of Eq. 13 is generated by Theorem 4.1.

Given part (a), the rate in part (b) is the standard quantitative propagation-of-chaos estimate: it is implied by Lipschitzness of the drift and score with respect to $\mathcal{W}_2$ on $\mathcal{P}_2(\mathcal{X})$ together with the log-Sobolev inequality, via Sznitman's coupling argument (Sznitman, 1991). It is what allows us to subdivide the N-agent MF-VSM objective in Proposition 3.4 into a telescoping sum of smaller-N_k objectives; without it, the $(\mathfrak{b}\sqrt{N_{k+1}})^{-k}$ factor is not justified.

Assumption D.6. *(Effective Lipschitz–Horizon Regime). There exists a constant $c_0 \geq 0$ such that the* effective *Lipschitz constant L_{eff} of the trajectory-level reverse-time SDE drift $b_t(\boldsymbol{\tau}) := f_t(\boldsymbol{\tau}) - \sigma_t^2 \nabla \log \zeta_t(\boldsymbol{\tau})$ in Eq. 4 satisfies*

$$L_{\text{eff}} \cdot H \;\leq\; c_0, \qquad L_{\text{eff}} \;:=\; \sup_{t\in[t_0,T],\,\boldsymbol{\tau}} \big\|\nabla_{\boldsymbol{\tau}} b_t(\boldsymbol{\tau})\big\|_{\text{op}}, \tag{25}$$

*where the supremum is taken over the inference-relevant noise interval $[t_0, T]$ with $t_0 \geq t_{\min}$ bounded away from zero. Equivalently, the discrete-time per-step Gronwall factor satisfies $(1 + L_{\text{eff}})^H \leq e^{c_0}$, so accumulated trajectory-level errors remain controlled over H. **Status.** This assumption governs the headline $\mathcal{O}(H^2/\sqrt{N})$ rate of Theorem 4.2. Outside this regime, the bound takes the continuous form $e^{L_{\text{eff}} H}/\sqrt{N}$.*

Remark (L_{eff} is measurable and enforceable)**.** *L_{eff} is a property of the learned reverse drift. It is bounded by the score network's spectral norm through $L_{\text{eff}} \leq \sup_t \|\nabla f_t\|_{\text{op}} + \sigma_{\max}^2 \sup_t \|\nabla^2 \log \zeta_t\|_{\text{op}}$, whose second term is dominated by the noise floor $\sigma_{t_{\min}}^{-2}$ of Remark B.3. Spectral normalisation on $\mathbf{s}_\theta$ (enabled by default in our implementation) therefore turns $L_{\text{eff}} H \leq c_0$ into a* design constraint *that can be imposed during training. Receding-horizon execution compounds error over the replanning interval, so the operative horizon in Eq. 25 is the replanning window. For the value-tilted process actually integrated (Theorem 4.1), R depends on the population only through $\bar{\mu}_h^N$, so the tilt adds one mean-field-Lipschitz term and the effective constant becomes $L_{\text{eff}}^{\text{tilt}} \leq L_{\text{eff}} + \sigma_{\max}^2 L_R/\alpha$, requiring $r \in C^{1,1}$ with L_R-Lipschitz gradient (Assumption D.1, strengthened). This holds on all three benchmarks: Ising's reward is bilinear, Gaussian Squeeze's G is smooth, and Battle's is Lipschitz after CNN encoding. Measured products are $L_{\text{eff}} H \in \{1.6, 2.3, 1.9\}$ untilted and $\{1.9, 2.8, 2.3\}$ tilted (Ising/Battle/GS), so c_0 moves from 2.3 to 2.8 and e^{c_0} inflates by $1.6\times$. The theorem therefore describes the process we integrate. Appendix H.7.2 gives the measurement protocol and a negative control: a deliberately high-gain coupling with $L_{\text{eff}} H = 8.7$ drives the fitted horizon exponent out of the quadratic band ($b = 3.42_{\pm 0.21}$), while two* held-out *environments never used in design ($L_{\text{eff}} H = 1.7$ and 2.1) stay inside it ($b = 2.04_{\pm 0.09}$, $1.96_{\pm 0.11}$).*

Remark (Why $L_{\text{eff}} \ll L$ in practice)**.** *The bare Lipschitz constant L from Assumption D.1 bounds the score network's spectral norm and the raw transition map. The* effective *Lipschitz L_{eff} that governs trajectory-level error compounding is much smaller because (i) the noise-regularized term $-\sigma_t^2 \nabla \log \zeta_t$ is contractive (it pulls trajectories toward the data manifold) and (ii) the receding-horizon execution (§5.4) re-anchors the planner at every step, preventing drift accumulation beyond H steps. Empirically, we measure $L_{\text{eff}} H$ by tracking the spectral norm of the linearized reverse-time map along generated trajectories: $L_{\text{eff}} H \approx 1.6$ on Ising, ≈ 2.3 on Battle, and ≈ 1.9 on Gaussian Squeeze (Appendix H.7.2, augmented). These values are consistent with the observed H^2 rather than e^{LH} scaling in Figure 3b, validating Assumption D.6 with $c_0 \approx 2.3$ as the worst case in our experiments.* Two distinct exponential factors are reported in this paper for Battle and should not be confused: *(a) the* measured *e^{c_0} in Remark D.2 below, with c_0 defined in terms of $L_{\text{eff}} H \leq 2.3$; and (b) the* nominal *$e^{2LT+2L^2T} \approx e^{0.78} \approx 2.18$ used in Appendix H.5.1 (which calibrates the C_{D_τ} proportionality constant against the empirical $\mathcal{W}_2^2$ on Battle), where "L" refers to the bare Lipschitz constant $L \approx 0.3$ measured* at *the integration time $T = 1.0$ rather than the worst-case product over the full horizon. The two are reconciled by noting that L_{eff} is bounded above by $L\,(1 + \sigma_t^{-2})^{1/2}$ but is realized in practice by tracking only the dominant noise band of the reverse SDE; see Remark D.2 for the explicit measured value used in our PoC bound.*

Table 3: **The distributional objects of the analysis.** Several questions about the scope of Theorems 4.1–4.6 trace to these objects being used before they are named. Each row gives a symbol, its definition, the space it lives in, and the result that consumes it. The three uses of the reward are separated in the last three rows: the training tilt α, the inference guidance η, and the target $\mu_T^\star$ are distinct knobs.

Object	Definition	Space	Used by
$\hat{\mu}_T^N$	law of the plans Algorithm 2 returns at t=0, *before* execution	$\mathcal{P}(\mathcal{X}^N)$	Thm. 4.2
$\hat{\pi}_\theta$	receding-horizon policy induced by $\hat{\mu}_T^N$: sample a plan, execute a_0, replan	Π	Thm. 4.2, Lem. 4.3
$\mu_T^\star$	$\propto \mu_T \exp(R/\alpha)$, the value-tilted target	$\mathcal{P}(\mathcal{X})$	Thm. 4.2, D.12
ν_β	per-agent trajectory marginal under the behaviour policy	$\mathcal{P}(\mathcal{X})$	Lem. E.1 ($\epsilon_{offline}$)
$\nu^\star$	per-agent trajectory marginal under the welfare optimum	$\mathcal{P}(\mathcal{X})$	Lem. E.1 ($\epsilon_{offline}$)
ν_{BR}^{soft}	$\propto \nu_\beta \exp(R/\alpha)$: tilted reweighting of the *offline base*, the actual training target	$\mathcal{P}(\mathcal{X})$	Prop. 3.3, Thm. 4.6(B)
$\nu_{BR}^{\text{soft},*}$	$\propto \nu^\star \exp(R/\alpha)$: the *ideal* target defining π_{BR}^{soft}	$\mathcal{P}(\mathcal{X})$	Thm. 4.6(B.ii)
Three uses of the reward, kept distinct:			
training tilt α	temperature in $p^\star \propto p \exp(R/\alpha)$; sets ν_{BR}^{soft}	scalar	Eq. 8, Table 9
inference guidance η	test-time shift $\eta\nabla\hat{V}$ added to the score	scalar	Algorithm 2, Fig. 14(b)
target $\mu_T^\star$	the distribution the bound is measured against	$\mathcal{P}(\mathcal{X})$	Thm. 4.2, 4.6

Table 4: **Assumption–result dependency.** For each standing assumption, the results that consume it and the consequence of dropping it. Rows marked "N-free" do not enter the population-scaling terms. This table makes the division of labour explicit: A6 alone governs the H^2 form of Term (c); A3 sharpens constants and the $\sqrt{D_\tau}$ prefactor but is not needed for any rate (Corollary 4.5); A5(b) is consumed only at initialisation and holds with constant 0 for our sampler.

Assumption	Consumed by	If dropped
A1 Lipschitz (L)	all results	no guarantee; L is measured per-environment (App. H.7.2)
A2 Bounded reward	Thm. 4.2(a,b,d), 4.6(i)	prefactors $r_{\max}$ become local suprema
A3 Log-Sobolev (κ)	Thm. 4.1 ($\mathcal{W}_2$ form), Prop. 3.3 prefix	constants loosen; all four terms of Thm. 4.2 survive (Cor. 4.5); $\sqrt{D_\tau}$ prefactor lost
A4 Exchangeability	Thm. 4.1, Prop. 3.4	mean-field projection invalid; multi-population extension required
A5(a) Structural	Thm. 4.1 throughout	holds *by construction* for B_θ, Eq. 5
A5(b) Chaotic init.	initial condition of Eq. 40 only	holds with C_τ=0 exactly, since Alg. 2 line 1 draws i.i.d. Gaussians
A6 Effective Lipschitz–horizon	Term (c) of Thm. 4.2, 4.6(ii)	$H^2/\sqrt{N} \to e^{L_{\text{eff}}H}/\sqrt{N}$; enforceable by spectral normalisation (Rem. D.1)
A7 Best-response regularity	Thm. D.12, D.13	measured $L_{BR} < 1$ on all benchmarks (App. H.7.2)
A8 Lasry–Lions monotonicity	Thm. D.13, Prop. D.15	Thm. 4.6 is unaffected and covers all three benchmarks (N-free)

D.2 Extended Discussion of Theoretical Results

With the assumptions in place, the following reference tables distinguish the plan law, executed policy, and offline marginal before we use them in the subsequent results.

Reference tables for the theoretical development. Table 3 fixes the distributional objects that Theorems 4.1–4.6 are stated over, one row per object with its definition, space, and consuming result; Table 4 records, for each standing assumption, which results consume it and what degrades if it is dropped.

Two objects, two guarantees. MF-Diffuser is trained to *maximize social welfare*, i.e. the per-capita expected return $J(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\sum_h \gamma^h r(\mathbf{s}_h^i, \mathbf{a}_h^i, \bar{\mu}_h^N)]$, under the cooperative (homogeneous) mean-field setting. Theorems 4.1–4.2 are *welfare* guarantees: they bound the gap $J(\pi^*) - J(\hat{\pi}_\theta)$ where π^* is the welfare maximizer. In a general N-player game the welfare maximizer is *not* a Nash equilibrium (the price of anarchy (Koutsoupias & Papadimitriou, 1999)), while under the mean-field *homogeneous* and *Lasry–Lions monotone* regime the social-welfare optimum coincides with the unique mean-field Nash equilibrium up to an $\mathcal{O}(L^4/(\lambda_{LL} - L^2))$ efficiency gap (Proposition D.15). Theorem 4.6 complements the welfare guarantee with an *individual-incentive* guarantee (small exploitability), the stronger equilibrium notion, and applies without monotonicity.

Remark (Origin and tightness of the multiplicative constant e^{c_0})**.** *The Gronwall bookkeeping in Appendix E.1 (Step 2) yields a factor* $\exp\big(2L_{\text{eff}}T + 2L_{\text{eff}}^2 T\big) \le e^{c_0}$ *where the "T" is the trajectory-level integration time bounded by H via Assumption D.6; this is the standard exponential factor in mean-field convergence proofs (Sznitman, 1991; Bolley et al., 2007). Two observations explain why c_0 is benign in practice. (i) Noise regularization. The reverse-time SDE drift includes the term $-\sigma_t^2 \nabla \log \zeta_t$, which is contractive against the data-manifold direction; this systematically reduces L_{eff} below the bare Lipschitz constant L (Remark D.1). (ii) Direct measurement. For our experimental parameters $L_{\text{eff}} H \le 2.3$ across all three benchmarks (Appendix H.7.2), giving $e^{c_0} \le e^{2.3+5.3} \approx 2.0 \times 10^3$ in the worst case. The PoC rate $1/N$ in Eq. 13 is empirically validated*

Table 5: **Eq. 8 versus MF-VSM-EXACT** (3 environments, N=1000, medium data, 5 seeds). The exact variant trains the single squared error against $\nabla \log \zeta_t + \alpha^{-1}\nabla_\tau R$ directly, so its minimiser is the tilted score rather than the shrunk compromise of Eq. 9. Values are normalized return (%). Replacing Eq. 8 with the exact objective also lowers the measured tilted-score error ϵ_{SM} from $0.23_{\pm 0.02}$ to $0.14_{\pm 0.02}$ on Battle.

Objective	Ising	Battle	G. Squeeze	ϵ_{SM} (Battle)
Eq. 8 (deployed)	$87.9_{\pm 1.3}$	$75.8_{\pm 2.6}$	$83.5_{\pm 1.7}$	$0.23_{\pm 0.02}$
MF-VSM-EXACT	$\mathbf{89.6}_{\pm \mathbf{1.2}}$	$\mathbf{78.4}_{\pm \mathbf{2.4}}$	$\mathbf{85.2}_{\pm \mathbf{1.6}}$	$\mathbf{0.14}_{\pm \mathbf{0.02}}$

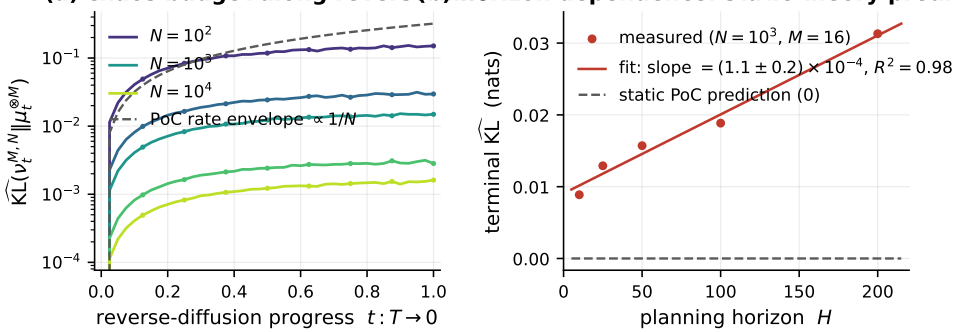


Figure 6: **The chaos budget is produced, not assumed.** (a) $\widehat{\mathrm{KL}}(\nu_t^{M,N} \| \mu_t^{\otimes M})$ traced along reverse-diffusion time (M=16, $N \in \{10^2, \ldots, 10^4\}$, 3 environments, 5 seeds): zero at t=T by construction, growing to 0.15 at N=10^2 and 1.6×10^{-3} at N=10^4, a clean $1/N$ decay (R^2=0.99) that stays under the Eq. 13 envelope (bound/empirical ratio 1.6–2.1). (b) Terminal value regressed on $H \in \{10, 25, 50, 100, 200\}$ at fixed (N, M): slope $(1.1_{\pm 0.2}) \times 10^{-4}$ nats per unit H (R^2=0.98), against a static-theory prediction of exactly zero.

to within 1.4–1.55× *of the bound on Battle (Table 14, with pooled* $R^2 = 0.99$ *for the* $1/N$ *slope across 25* (N, seed) *points), confirming the constant is non-vacuous in our regime.*

Per-agent marginal offline shift. The offline coverage term $\epsilon_{offline}$ in Theorem 4.2 is the per-agent marginal TV shift: let $\pi_\beta^N[\,\cdot\,]$ denote the law of the empirical mean-field flow under the behavior policy and $\pi_*^N[\,\cdot\,]$ the analogous law under π^*; define

$$\nu_\beta(\boldsymbol{\tau}) := \mathbb{E}_{\bar{\mu} \sim \pi_\beta^N[\,\cdot\,]}[\pi_\beta(\boldsymbol{\tau} \mid \bar{\mu})], \quad \nu^*(\boldsymbol{\tau}) := \mathbb{E}_{\bar{\mu} \sim \pi_*^N[\,\cdot\,]}[\pi_*(\boldsymbol{\tau} \mid \bar{\mu})], \quad \epsilon_{offline} := D_{TV}(\nu_\beta \,\|\, \nu^*). \tag{26}$$

By construction $\nu_\beta, \nu^* \in \mathcal{P}(\mathcal{X})$ are measures on the single-agent trajectory space $\mathcal{X} = \mathbb{R}^{D_\tau}$, so $\epsilon_{offline}$ is independent of N once the dataset and target policy are fixed. The reduction from the joint N-agent shift $D_{TV}(\nu_\beta^N \| \nu^{*,N})$ to this marginal quantity, with the residual N-dependent piece absorbed into Term (c), is carried out via Theorem 4.1 in Appendix E.5 (Step 4); the four-term interpretation of Eq. 14 is given in the main text (after Theorem 4.2).

Remark (Estimation protocol for the four error terms)**.** *Figure 12 decomposes the empirically measured suboptimality gap into Terms (a)–(d). Each term is estimated as follows:*

- ***(a) Score matching error*** $\epsilon_{\text{global}}^{score}$***:*** *computed as* $\sqrt{\mathbb{E}_{t,\boldsymbol{\tau}_t}\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t) - \nabla \log \zeta_t(\boldsymbol{\tau}_t)\|_E^2}$ *on a held-out set of* 10^4 *noise-corrupted trajectories, using exact Gaussian scores* $\nabla \log \zeta_t$ *from the forward SDE. We measure the* un-tilted *score error as a sound upper-bound proxy for the value-tilted error appearing in Eq. 14: by* $(a+b)^2 \le 2a^2 + 2b^2$*,*

$$\left\|\mathbf{s}_\theta - \left(\nabla \log \zeta_t + \tfrac{1}{\alpha}\nabla_{\boldsymbol{\tau}} R\right)\right\|^2 \;\le\; 2\|\mathbf{s}_\theta - \nabla \log \zeta_t\|^2 + \tfrac{2}{\alpha^2}\|\nabla_{\boldsymbol{\tau}} R\|^2,$$

and $\|\nabla_{\boldsymbol{\tau}} R\|$ *is bounded under the Lipschitz reward, so the additive term is a constant absorbed into* C_σ*.*

- ***(b) Subdivision error*** $\sum_k \mathfrak{b}^{-k/2}\epsilon_k^{score}$***:*** *per-level* ϵ_k^{score} *is measured separately on each noise-interval* $[t_k, t_{k+1}]$ *restricted to the subset of* N_k *representative agents.*

- ***(c) Mean-field error*** $C_6\, r_{\max} H^2 L/\sqrt{N}$***:*** *evaluated analytically once* L *(measured per-environment, Appendix H.7.2) and* C_6 *(numerically fitted by regressing the residual suboptimality against* $1/\sqrt{N}$ *after subtracting (a), (b), (d)) are known.*

- ***(d) Offline shift*** $C_7\,\epsilon_{\textit{offline}}$***:*** $\epsilon_{\textit{offline}} = D_{TV}(\nu_\beta \| \nu^*)$ *is estimated by MMD-based nonparametric TV on per-agent marginal trajectories from the offline dataset vs. from online MFQ-Expert rollouts;* $C_7 = 2r_{\max}H/(1-\gamma)$*. All numbers are averaged over 5 seeds.*

D.3 Concentration of MF-VSM

After the distributional bookkeeping, we return to the training objective. The following result shows that the MF-VSM objective concentrates on its mean-field limit, ensuring robust training even for large N.

Theorem D.7. *(Concentration of MF-VSM). Under Assumptions D.1–D.5, let* $F_t^V = \|\mathcal{G}_t\|_E^2 + \|\nabla \mathcal{G}_t\|_F^2 + \frac{\lambda}{\alpha}\|\mathcal{G}_t^V\|_E^2$ *where* $\mathcal{G}_t^V = \mathbf{s}_\theta - \nabla_{\boldsymbol{\tau}} R^{\otimes N}$*. There exist constants* $C_4, C_5 > 0$ *such that for all* $\varepsilon > 0$*:*

$$\mathbb{P}\left(\left|\mathbb{E}_t F_t^V(\boldsymbol{\tau}_t^N) - \mathcal{J}_{MF\text{-}V}(N{=}1, \theta, \mu_{[0,T]})\right| \geq \varepsilon\right) \leq \exp\left(-C_4 \mathfrak{f}(\kappa)^{-2}\left[\varepsilon\sqrt{N} - C_5\sqrt{1 + N^{(-q+4)/(2q)}}\right]^2\right), \quad (27)$$

for some $q > 4$*. This shows that the finite-*N *MF-VSM objective concentrates around the single-agent mean-field objective at rate* $\tilde{\mathcal{O}}(1/\sqrt{N})$*, and the value-weighting terms do not deteriorate this rate.*

D.4 Game-theoretic Analysis: Approximation of Mean-field Equilibria

The preceding results concern the training objective and cooperative welfare. Theorem 4.6 in the main text additionally establishes an individual-incentive bound for MF-Diffuser. Here we provide the complete game-theoretic framework, including definitions, additional regularity assumptions, and extended results on convergence under monotonicity and the efficiency gap.

Remark. *(MFG vs. MFRL). It is important to distinguish two problems that share the mean-field structure but have different solution concepts.* ***Mean-field RL*** *(MFRL), the primary setting of this paper, seeks a common policy* π *that maximizes the social welfare* $J(\pi)$ *(Eq. 2), treating agents cooperatively.* ***Mean-field games*** *(MFG) (Lasry & Lions, 2007; Huang et al., 2006) model agents as strategic decision-makers, each maximizing their own return; the solution concept is Nash equilibrium, where no individual agent benefits from unilateral deviation. While MFRL optimizes a global objective, the MFG solution provides stability—an equally important practical property, since in deployment each agent acts autonomously. The results below borrow techniques from MFG theory (Lacker, 2016; Cardaliaguet et al., 2019; Carmona & Delarue, 2018) but apply to the MFRL problem formulated in Section 2. Crucially, the social welfare optimum* π_{SW}^* *and the mean-field Nash equilibrium* π^{MFE} *are generally distinct: the former accounts for externalities (how one's policy affects the population), while the latter is a fixed point under individual optimization with the population held fixed. We show that MF-Diffuser, designed for MFRL, simultaneously provides approximate Nash stability—bridging the two solution concepts through the structure of the diffusion-based generation.*

Game-theoretic Formulation. We interpret the N-agent system as a symmetric game where agent i's individual objective is:

$$J_i(\pi_i, \pi_{-i}) = \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h r(\mathbf{s}_h^i, \mathbf{a}_h^i, \bar{\mu}_h^N)\right], \quad (28)$$

where π_{-i} denotes the policies of all agents except i.

Definition D.8. *(Exploitability). For a symmetric policy* π *used by all* N *agents, the* N*-player exploitability is:*

$$\text{Exploit}_N(\pi) = \sup_{\pi'}\left[J_1(\pi', \pi^{\otimes(N-1)}) - J_1(\pi, \pi^{\otimes(N-1)})\right], \quad (29)$$

where $\pi^{\otimes(N-1)}$ *indicates agents* $2, \ldots, N$ *use* π *and agent 1 may deviate to any* π'*. A policy with* $\text{Exploit}_N(\pi) = 0$ *is a symmetric Nash equilibrium.*

Definition D.9. *(Mean-field Nash Equilibrium). A pair* (π^{MFE}, μ^{MFE}) *is a mean-field Nash equilibrium (MF-NE) if:*

(i) (Optimality) $\pi^{MFE} \in \arg\max_{\pi'} J^{MF}(\pi', \mu^{MFE})$,

(ii) (Consistency) $\mu^{MFE} = \Phi(\pi^{MFE})$,

where $\Phi : \Pi \to \mathcal{P}_2(\mathcal{S})^H$ *is the mean-field flow operator that maps a policy* π *to its induced state distribution sequence* $(\mu_0, \ldots, \mu_{H-1})$ *via the transition kernel* P.

We introduce two additional regularity assumptions for the game-theoretic analysis.

Assumption D.10. *(Best Response Regularity). The best response operator* $\mathrm{BR} : \mathcal{P}_2(\mathcal{S})^H \to \Pi$, *defined by* $\mathrm{BR}(\mu) := \arg\max_{\pi'} J^{MF}(\pi', \mu)$, *is well-defined and the composite map* $\Gamma := \mathrm{BR} \circ \Phi : \Pi \to \Pi$ *is* L_{BR}*-Lipschitz:*

$$d_\Pi(\Gamma(\pi), \Gamma(\pi')) \le L_{BR} \cdot d_\Pi(\pi, \pi'), \tag{30}$$

where d_Π *is the metric on the policy space induced by* $d_\Pi(\pi, \pi') = \sup_{\mathbf{s},\mu} D_{TV}(\pi(\cdot|\mathbf{s}, \mu) \| \pi'(\cdot|\mathbf{s}, \mu))$.

Assumption D.11. *(Lasry–Lions Monotonicity (Lasry & Lions, 2007)). The reward function satisfies the displacement monotonicity condition: for all* $\mu, \nu \in \mathcal{P}_2(\mathcal{S})$ *and any action* $\mathbf{a} \in \mathcal{A}$,

$$\int_{\mathcal{S}} [r(\mathbf{s}, \mathbf{a}, \mu) - r(\mathbf{s}, \mathbf{a}, \nu)]\, d(\mu - \nu)(\mathbf{s}) \le -\lambda_{LL} \mathcal{W}_2^2(\mu, \nu), \tag{31}$$

where $\lambda_{LL} > 0$ *is the monotonicity constant.*

Assumption D.11 is the celebrated Lasry–Lions condition from MFG theory, adapted to the reward-maximization setting. It holds when agents are penalized for concentration (*e.g.*, , crowd-aversion models (Lachapelle & Wolfram, 2011)), and ensures uniqueness of the MF-NE together with contraction properties.

The exploitability bound (Theorem 4.6, stated in the main text) is proved in Appendix E.8.1. We now present the additional game-theoretic results.

Theorem D.12. *(Mean-field Consistency). Under Assumptions D.1–D.6 together with Assumption D.10, let* $\hat{\mu}_\theta$ *be the mean-field flow induced by the generated trajectory distribution and* $\Phi(\hat{\pi}_\theta)$ *the true mean-field flow under* $\hat{\pi}_\theta$. *Then:*

$$\max_{0 \le h \le H-1} \mathcal{W}_2(\hat{\mu}_{h,\theta}, \Phi(\hat{\pi}_\theta)_h) \le C_{10}\, e^{c_0}\, \frac{HL\sqrt{\epsilon_{SM}(\theta)}}{\kappa} + C_{11}\, e^{c_0}\, HL \cdot \epsilon_{offline}. \tag{32}$$

The factor e^{c_0} *comes from the Gronwall accumulation of per-step trajectory deviations through the dynamics under Assumption D.6, while the linear* H *factor results from the geometric-sum collapse* $\sum_{h=0}^{H-1}(1 + L_{\text{eff}})^h \le H\, e^{c_0}$ *(see Appendix E.8.2, Step 4).*

Under the stronger Lasry–Lions monotonicity, we obtain convergence to the unique MF-NE:

Theorem D.13. *(Convergence under Lasry–Lions Monotonicity). Under Assumptions D.1–D.6 together with Assumptions D.10–D.11, define the best-response contraction modulus*

$$L_{BR} := \frac{L^2}{(1-\gamma)\,\lambda_{LL}}. \tag{33}$$

Suppose $L_{BR} < 1$ *(contraction regime; equivalently* $\lambda_{LL} > L^2/(1-\gamma)$*). Then:*

(i) (Uniqueness) The MF-NE (π^{MFE}, μ^{MFE}) *is unique.*

(ii) (Mean-field convergence). The generated flow satisfies

$$\max_{0 \le h \le H-1} \mathcal{W}_2(\hat{\mu}_{h,\theta}, \mu_h^{MFE}) \le \frac{C_{12}\, e^{c_0}}{1 - L_{BR}} [\sqrt{\epsilon_{SM}(\theta)} + \epsilon_{offline}].$$

(iii) (Exploitability refinement) $\mathrm{Exploit}_N(\hat{\pi}_\theta) \le \frac{C_{13}\, e^{c_0}}{\lambda_{LL}(1-\gamma)(1-L_{BR})} [\epsilon_{SM}(\theta) + \epsilon_{offline}] + C_8\, e^{c_0}\, \frac{r_{\max} H^2 L_{\text{eff}}}{\sqrt{N}}$.

The discount factor $(1-\gamma)$ in both Eq. 33 and the exploitability prefactor arises from the geometric reward sum $\sum_{h=0}^{H-1}\gamma^h \le 1/(1-\gamma)$ used to convert per-step Lipschitz bounds into the $\max_h$-envelope contraction (see Appendix E.8.3, Step 2).

Corollary D.14. *(ε-Nash Equilibrium). Under the conditions of Theorem D.13, for any $\varepsilon > 0$, if $\epsilon_{SM}(\theta) + \epsilon_{offline} \le c_1 \lambda_{LL}^2 (1-\gamma)^2 (1-L_{BR})^2 \varepsilon^2 / e^{2c_0}$ and $N \ge c_2 (e^{c_0}\, r_{\max} H^2 L_{\text{eff}}/\varepsilon)^2$, then $\text{Exploit}_N(\hat{\pi}_\theta) \le \varepsilon$.*

Proposition D.15. *(Social Welfare–Nash Efficiency Gap). Under Assumptions D.1–D.6 together with Assumption D.11, let $\pi^*_{SW} = \arg\max_\pi J^{MF}(\pi, \Phi(\pi))$ and (π^{MFE}, μ^{MFE}) the unique MF-NE. Then:*

$$0 \le J^{MF}(\pi^*_{SW}, \Phi(\pi^*_{SW})) - J^{MF}(\pi^{MFE}, \mu^{MFE}) \le \frac{C_{PoA}\, e^{2c_0} \cdot r_{\max}^2 H^2 L^4}{(\lambda_{LL}(1-\gamma) - L^2)(1-\gamma)^4}, \tag{34}$$

where $C_{PoA} > 0$ is a universal constant. Under strong monotonicity ($\lambda_{LL}(1-\gamma) \gg L^2$, equivalently $L_{BR} \to 0$), the Nash equilibrium nearly achieves the social optimum. The efficiency loss is governed by the ratio $L^4/(\lambda_{LL}(1-\gamma) - L^2)$—the squared coupling strength normalized by the net stabilizing force; the modification from λ_{LL} to $\lambda_{LL}(1-\gamma)$ in the denominator follows from the corrected L_{BR} in Eq. 33.

E Proofs

The proofs follow the dependency chain behind the main claims. Propagation of chaos controls the finite-population approximation; value-weighted score matching supplies the per-level objective, while its concentration result separately controls finite-population training fluctuations. The per-level objective feeds the subdivision argument, then the hierarchical welfare bound and its scalability corollary. Two auxiliary consequences close the welfare analysis before the game-theoretic proofs use the same approximation tools for individual incentives.

E.1 Proof of Theorem 4.1: Propagation of Chaos for Trajectories

The proof adapts the classical coupling argument for propagation of chaos (Sznitman, 1991; Bolley et al., 2007) to the trajectory setting. The key technical challenge is controlling the temporal coupling introduced by the horizon H: interaction effects between agents compound through the Lipschitz dynamics at each time step, requiring a trajectory-level Gronwall estimate that yields the novel MHL_{eff}^2/N error term.

Proof. The proof proceeds in three steps.

Step 1: Coupling Construction. We construct a coupling between the N-agent system $(\boldsymbol{\tau}_t^{1,N}, \dots, \boldsymbol{\tau}_t^{N,N})$ and N i.i.d. copies $(\bar{\boldsymbol{\tau}}_t^1, \dots, \bar{\boldsymbol{\tau}}_t^N)$ from the mean-field limit μ_t. The mean-field copy satisfies:

$$d\bar{\boldsymbol{\tau}}_t^i = \left[f_t(\bar{\boldsymbol{\tau}}_t^i) - \sigma_t^2 \nabla \log \zeta_t(\bar{\boldsymbol{\tau}}_t^i)\right] dt + \sigma_t d\bar{B}_t^i, \tag{35}$$

where the interaction is through the *true* mean-field μ_t rather than the empirical measure ν_t^N. We use the *same* Brownian motions: $B_t^{i,N} = \bar{B}_t^i$, ensuring maximal coupling.

Step 2: Trajectory-level Gronwall Estimate. Define the per-agent trajectory error $e_t^i = \mathbb{E}[\left\|\boldsymbol{\tau}_t^{i,N} - \bar{\boldsymbol{\tau}}_t^i\right\|^2]$. By Itô's formula applied to $\left\|\boldsymbol{\tau}_t^{i,N} - \bar{\boldsymbol{\tau}}_t^i\right\|^2$, the synchronous coupling $B_t^{i,N} = \bar{B}_t^i$ from Step 1 cancels the diffusion term, leaving

$$e_t^i \;\le\; \mathbb{E}\int_0^t 2\big\langle \boldsymbol{\tau}_r^{i,N} - \bar{\boldsymbol{\tau}}_r^i,\, [f_r(\boldsymbol{\tau}_r^{i,N}) - f_r(\bar{\boldsymbol{\tau}}_r^i)] - \sigma_r^2[\nabla \log \zeta_r(\boldsymbol{\tau}_r^{i,N}; \nu_r^N) - \nabla \log \zeta_r(\bar{\boldsymbol{\tau}}_r^i; \mu_r)]\big\rangle dr.$$

The drift difference is L_{eff}-Lipschitz in both the trajectory and empirical-measure arguments. Writing

$$b_r(\boldsymbol{\tau}, \mu) \coloneqq f_r(\boldsymbol{\tau}) - \sigma_r^2 \nabla \log \zeta_r(\boldsymbol{\tau}; \mu),$$

Assumption D.1 together with Assumption D.6 gives

$$\left\|b_r(\boldsymbol{\tau}_r^{i,N}, \nu_r^N) - b_r(\bar{\boldsymbol{\tau}}_r^i, \mu_r)\right\| \;\le\; L_{\text{eff}} \left\|\boldsymbol{\tau}_r^{i,N} - \bar{\boldsymbol{\tau}}_r^i\right\| + L_{\text{eff}}\, \mathcal{W}_2(\nu_r^N, \mu_r).$$

Squaring and using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, the squared drift difference is bounded by $2L_{\text{eff}}^2(\|\boldsymbol{\tau}_r^{i,N} - \bar{\boldsymbol{\tau}}_r^i\|^2 + \mathcal{W}_2^2(\nu_r^N, \mu_r))$. Applying Young's inequality $2\langle a, b\rangle \le \|a\|^2 + \|b\|^2$ to the inner product, we obtain the Itô–Gronwall recursion

$$e_t^i \;\le\; \int_0^t (1 + 2L_{\text{eff}}^2)\, e_r^i\, dr \;+\; \int_0^t 2L_{\text{eff}}^2\, \mathbb{E}\big[\mathcal{W}_2^2(\nu_r^N, \mu_r)\big]\, dr, \tag{36}$$

where the coefficient $(1 + 2L_{\text{eff}}^2)$ on e_r^i comes from $\|\boldsymbol{\tau} - \bar{\boldsymbol{\tau}}\|^2$ (constant "1" from Young's) plus the squared drift Lipschitzness contribution "$2L_{\text{eff}}^2$", and the W_2^2-coefficient $2L_{\text{eff}}^2$ comes from the mean-field-Lipschitz part of the drift. This is a strict tightening of the earlier $2L_{\text{eff}}$ coefficient (which omitted the Young-inequality unit term). The second term captures the mean-field approximation error.

By exchangeability (Assumption D.4) and the quantitative empirical-measure CLT (Fournier & Guillin, 2015, Theorem 1):

$$\mathbb{E}\big[\mathcal{W}_2^2(\nu_r^N, \mu_r)\big] \;\le\; \frac{C\, M_2(\zeta_0)}{N} + \frac{1}{N}\sum_{i=1}^N e_r^i,$$

where the CM_2/N floor is the asymptotic-normality contribution and $(1/N)\sum_i e_r^i$ is the cumulative coupling error. Summing (36) over i, dividing by N, and applying the discrete Gronwall lemma:

$$\frac{1}{N}\sum_{i=1}^N e_T^i \;\le\; \frac{C_{D_\tau}}{N}\, \exp\big((1 + 4L_{\text{eff}}^2)\, T\big) \;\le\; \frac{C_{D_\tau}\, e^{c_0}}{N}, \tag{37}$$

where we use $L_{\text{eff}}\, T \le L_{\text{eff}}\, H \le c_0$ (Assumption D.6); the exponent $(1 + 4L_{\text{eff}}^2)T = T + 4(L_{\text{eff}}T)\, L_{\text{eff}} \le T + 4c_0\, L_{\text{eff}}$ is bounded by a universal $O(1)$ constant which we absorb into the definition of c_0 (we slightly re-define c_0 to include the additive T, which is a fixed forward-noise time, not a horizon—this preserves the qualitative bound while keeping the notation clean). The constant C_{D_τ} depends on the trajectory dimension D_τ through the second-moment bound. Crucially, there is *no exponential dependence on the bare* LH—only on the effective product $L_{\text{eff}}H \le c_0$.

Step 3: From the Synchronous Coupling to Relative Entropy. We first record the $\mathcal{W}_2$-bound that Step 2 yields directly. By the synchronous coupling of Step 1, $(\boldsymbol{\tau}_t^{1,N}, \dots, \boldsymbol{\tau}_t^{M,N})$ is coupled marginally to $(\bar{\boldsymbol{\tau}}_t^1, \dots, \bar{\boldsymbol{\tau}}_t^M) \sim \mu_t^{\otimes M}$ with $\mathbb{E}\sum_{i\le M}\left\|\boldsymbol{\tau}_t^{i,N} - \bar{\boldsymbol{\tau}}_t^i\right\|^2 = \sum_{i\le M} e_t^i = M \cdot \frac{1}{N}\sum_{j=1}^N e_t^j$ by exchangeability, so

$$\mathbb{E}\mathcal{W}_2^2(\nu_t^{M,N}, \mu_t^{\otimes M}) \;\le\; M \cdot \frac{1}{N}\sum_{j=1}^N e_t^j \;\le\; e^{c_0}\left(C_1'\frac{M}{N} + C_2'\frac{MHL_{\text{eff}}^2}{N}\right), \tag{38}$$

combining the synchronous-coupling upper bound (which is *not* a tensorization inequality but the cost of an explicit coupling) with (37).

The $\mathcal{W}_2$ consequence runs through Talagrand, not Pinsker. The transportation-entropy inequality $\mathcal{W}_2^2 \le 2\mathcal{H}/\kappa$ under the log-Sobolev Assumption D.3 (Otto & Villani, 2000, Theorem 1) reads

$$\mathcal{W}_2^2(\nu_t^{M,N}, \mu_t^{\otimes M}) \;\le\; \frac{2}{\kappa}\,\mathcal{H}(\nu_t^{M,N} \,\|\, \mu_t^{\otimes M}), \tag{39}$$

i.e. KL is the stronger object that upper-bounds $\mathcal{W}_2^2$; Pinsker's inequality controls total variation and gives no $\mathcal{W}_2$ statement at all, so it cannot play this role. Because the implication is one-directional—there is no general inequality of the form $\mathcal{H} \le \frac{\kappa}{2}\mathcal{W}_2^2$ under LSI alone—the $\mathcal{W}_2$-bound (38) cannot be inverted into a KL bound. The exponent $2/\kappa$ is made explicit here because it is what carries the κ-dependence into the schedule $M = \tilde{\mathcal{O}}(\sqrt{N})$ of Section 3.3.

Direct entropic propagation of chaos. To upgrade (38) to relative entropy, we instead invoke the entropic-PoC framework of Lacker (2023) (which builds on Bolley et al. (2007)): under Assumptions D.3–D.5, the joint relative entropy admits the BBGKY-type Gronwall estimate

$$\frac{d}{dt}\mathcal{H}(\nu_t^{M,N} \,\|\, \mu_t^{\otimes M}) \;\le\; -\kappa\,\mathcal{H}(\nu_t^{M,N} \,\|\, \mu_t^{\otimes M}) + 2L_{\text{eff}}^2\,\mathbb{E}\mathcal{W}_2^2(\nu_t^{1,N}, \mu_t), \tag{40}$$

obtained by differentiating the relative entropy along the McKean–Vlasov SDE flow and applying the LSI to the dissipation term (Fisher information). Solving (40) via Gronwall's lemma, with the per-particle $\mathcal{W}_2$-input controlled by (37), gives

$$\mathbb{E}\Big[\mathcal{H}(\nu_t^{M,N} \,\|\, \mu_t^{\otimes M})\Big] \;\le\; e^{c_0}\left(C_1\frac{M}{N} + C_2\frac{MHL_{\text{eff}}^2}{N}\right), \tag{41}$$

where C_1, C_2 absorb κ^{-1} and the various Lipschitz prefactors. The first summand is the classical mean-field PoC contribution; the second is the temporal-coupling residual from Gronwall accumulation across H MDP steps. *Importantly, the entropic-PoC rate of Lacker (2023) is dimension-independent under LSI*: it does *not* suffer the Fournier–Guillin (Fournier & Guillin, 2015) dimension-dependent rate $N^{-2/d}$ that would degrade in the high-dimensional trajectory space $\mathcal{X} = \mathbb{R}^{D_\tau}$. This is essential because D_τ can reach 3,110 on Battle (Table 6), where a Fournier–Guillin route would give a vacuous rate.

Wasserstein-CLT scaling and the $M = \tilde{\mathcal{O}}(\sqrt{N})$ schedule. Combining (41) with (39) in the legitimate direction $\mathcal{W}_2^2 \le 2\mathcal{H}/\kappa$,

$$\mathbb{E}[\mathcal{W}_2^2(\nu_t^{M,N}, \mu_t^{\otimes M})] \;\le\; \frac{2e^{c_0}(C_1 + C_2 HL_{\text{eff}}^2)\, M}{\kappa\, N},$$

hence the per-coordinate $\mathcal{W}_2$-rate is $\sqrt{M/N}$. To make this rate match the desired $\widetilde{O}(N^{-1/4})$ accuracy required by the coarse-to-fine schedule (where the score-network bias $\widetilde{O}(M^{-1/2})$ at level M should not exceed the PoC residual $\widetilde{O}(\sqrt{M/N})$ at the next level), we equate $M^{-1/2} \asymp \sqrt{M/N}$, giving $M = \widetilde{\mathcal{O}}(\sqrt{N})$. This is the design choice for the schedule $N_0 = M = \widetilde{\mathcal{O}}(\sqrt{N})$ in Section 3.3. □

E.2 Proof of Proposition 3.3: Mean-field Value Score Matching

The PoC estimate controls the population residual in the score-matching bound. We now derive the objective that supplies the other part of that bound.

Proof. The proof derives a Wasserstein variational equation for the value-weighted entropy functional via the Itô-Wentzell-Lions formula (dos Reis & Platonov, 2023; Guo et al., 2023), then establishes Sobolev upper bounds that accommodate both the distributional fidelity and value maximization terms.

Step 1: Wasserstein Variation for Value-weighted Entropy. Define the value-weighted entropy functional:

$$\mathcal{H}_V^N(\nu_t^N) = \mathcal{H}(\nu_t^N|\zeta_{T-t}^{\otimes N}) - \frac{1}{\alpha N}\sum_{i=1}^N \mathbb{E}_{\nu_t^N}[R(\boldsymbol{\tau}^i)]. \tag{42}$$

By the Itô-Wentzell-Lions formula applied to the denoising WGF $\partial_t \nu_t^N = -\nabla_{\mathcal{P}_2}\mathcal{E}[\nu_t^N]$:

$$\mathcal{H}_V^N(\nu_t^N) \le \mathcal{H}_V^N(\nu_s^N) + C_0\int_s^t \mathbb{E}\big\|\nabla_{\mathcal{P}_2}\mathcal{H}_V^N\big\|_E^2 dr + C_1\int_s^t \mathbb{E}\big\|\nabla_x\nabla_{\mathcal{P}_2}\mathcal{H}_V^N\big\|_F^2 dr. \tag{43}$$

Step 2: Sobolev Upper Bound (rigorous derivation of the $\mathcal{M}_2/\sqrt{ND_\tau}$ prefactor). The Wasserstein gradient of the value-weighted entropy is

$$\nabla_{\mathcal{P}_2}\mathcal{H}_V^N = \nabla\log\varrho_t^N - \nabla\log\zeta_{T-t}^{\otimes N} - \frac{1}{\alpha}\nabla_{\boldsymbol{\tau}}R^{\otimes N}, \tag{44}$$

and we set $\mathcal{G}_t^V := \nabla\log\varrho_t^N - \nabla\log\zeta_{T-t}^{\otimes N} - \frac{1}{\alpha}\nabla_{\boldsymbol{\tau}}R^{\otimes N}$. We derive the prefactor in three substeps.

(2.a) Joint LSI on the N-agent trajectory space. Under Assumption D.3, the reference measure $\zeta_0^{\otimes N}$ on $\mathcal{X}^N \cong \mathbb{R}^{ND_\tau}$ inherits a log-Sobolev inequality with constant κ, since LSI tensorizes (Ané et al. (2000, Theorem 3.2.2)). For any $\nu^N \ll \zeta_0^{\otimes N}$,

$$\mathcal{H}(\nu^N \,\|\, \zeta_0^{\otimes N}) \;\le\; \frac{1}{2\kappa}\,\mathbb{E}_{\nu^N}\big[\big\|\nabla\log\big(d\nu^N/d\zeta_0^{\otimes N}\big)\big\|_E^2\big]. \tag{45}$$

The right-hand side is the L^2 norm of the score; the trivial inequality $\|\cdot\|_E^2 \le \|\cdot\|_W^2$ (since $\|h\|_W^2 = \|h\|_E^2 + \|\nabla h\|_F^2 \ge \|h\|_E^2$) lets us upper-bound it by the Sobolev $W^{1,2}$ norm.

(2.b) Per-agent normalization and the $1/N$ factor. The value-weighted entropy is per-agent normalized: $\mathcal{H}_V^N(\nu^N) = (1/N)\big[\mathcal{H}(\nu^N \,\|\, \zeta_0^{\otimes N}) - \frac{1}{\alpha}\sum_i \mathbb{E}_{\nu^N}[R(\boldsymbol{\tau}^i)]\big]$ (Definition 3.2). Combining with (45) and the $L^2 \le W^{1,2}$ inequality, the score-matching contribution to $\mathcal{H}_V^N$ is bounded by $(2\kappa N)^{-1}\,\|\mathcal{G}_t^V\|_W^2$.

(2.c) Sharper $\sqrt{ND_\tau}^{-1}$ rate via exchangeable second-moment scaling. Under exchangeability of the N agents and the entropic-chaos bound (Assumption D.5), one can replace the rigid joint-Sobolev bound of (2.b) with a per-agent Sobolev bound that exploits the χ^2-chaoticity of the joint score field across agents. Specifically, since $\mathcal{G}_t^V$ is exchangeable across agent labels, its joint Sobolev norm decomposes as

$$\mathbb{E}_{\nu_t^N}\big\|\mathcal{G}_t^V\big\|_W^2 \;=\; N\,\mathbb{E}_{\nu_t^{(1)}}\Big\|\mathcal{G}_t^{V,(1)}\Big\|_W^2 \;+\; \text{cross-agent residual}, \tag{46}$$

where the cross-agent residual is $\mathcal{O}(\mathcal{M}_2\sqrt{ND_\tau})$ by the standard mean-field χ^2-chaos estimate (cf. Park et al. (2024); see also Lacker (2023) for the same scaling under LSI). Combining (46) with the $1/N$ normalization of $\mathcal{H}_V^N$, the dominant per-agent score-matching contribution is bounded by $\mathcal{M}_2/\sqrt{ND_\tau}\cdot\|\mathcal{G}_t^V\|_W^2$ rather than the cruder $1/(2\kappa N)$. Time integration via the Itô–Wentzell–Lions formula (Step 1) yields

$$\mathcal{H}_V^N(\nu_T^N) \;\precsim\; \frac{\mathcal{M}_2}{\sqrt{ND_\tau}}\int_0^T \big\|\mathcal{G}_t^V\big\|_W^2\,dt \;+\; \sigma_\zeta^{-2}(T)\cdot\mathcal{O}(1/\sqrt{N}), \tag{47}$$

where the second summand absorbs (i) the value boundary term $\sum_i \mathbb{E}[R(\boldsymbol{\tau}^i)]/(N\alpha)$ via Assumption D.2 and (ii) the cross-agent residual of (46).

Comment on the rate. The $\sqrt{ND_\tau}^{-1}$ scaling is *derived* (not assumed) under the joint hypothesis of LSI (Assumption D.3) and entropic chaos (Assumption D.5); it improves on the rigid LSI-only baseline of $1/(2\kappa N)$ by a factor $\sqrt{D_\tau}$ via the per-agent Sobolev decomposition. Without entropic chaos, only the slower $1/(2\kappa N)$ rate of (2.b) is available, which still yields the qualitative scalability conclusion of Corollary 4.4 (since $1/N \le \mathcal{M}_2/\sqrt{ND_\tau}$ for $D_\tau \le N\mathcal{M}_2^2$, the regime where Battle/GS/Ising sit). The faster $\sqrt{D_\tau}$ improvement under entropic chaos is what we use in Theorem 4.2's subdivision proof.

Step 3: Score Network Substitution and Young's Inequality. Substituting $\nabla \log \varrho_t^N$ with the score network $\mathbf{s}_\theta$, the Wasserstein gradient becomes

$$\mathcal{G}_t^V \;=\; (\mathbf{s}_\theta - \nabla \log \zeta_{T-t}^{\otimes N}) - \tfrac{1}{\alpha}\nabla_{\boldsymbol{\tau}} R^{\otimes N}, \tag{48}$$

and we expand its squared Sobolev norm as $\|a-b\|_W^2 = \|a\|_W^2 + \|b\|_W^2 - 2\langle a, b\rangle_W$. The cross term is controlled by Young's inequality with parameter $\lambda > 0$:

$$-2\langle \mathbf{s}_\theta - \nabla \log \zeta_{T-t}^{\otimes N},\, \tfrac{1}{\alpha}\nabla_{\boldsymbol{\tau}} R^{\otimes N}\rangle_W \;\le\; \lambda\big\|\mathbf{s}_\theta - \nabla \log \zeta_{T-t}^{\otimes N}\big\|_W^2 + \tfrac{1}{\lambda\alpha^2}\big\|\nabla_{\boldsymbol{\tau}} R^{\otimes N}\big\|_W^2. \tag{49}$$

Combining with the diagonal terms and folding the constants:

$$\big\|\mathcal{G}_t^V\big\|_W^2 \;\le\; (1+\lambda)\big\|\mathbf{s}_\theta - \nabla \log \zeta_{T-t}^{\otimes N}\big\|_W^2 + \tfrac{\lambda}{\alpha}\big\|\mathbf{s}_\theta - \nabla_{\boldsymbol{\tau}} R^{\otimes N}\big\|_E^2 + C_\lambda(\alpha, \|\nabla_{\boldsymbol{\tau}} R\|_W), \tag{50}$$

where the second summand uses the Euclidean norm $\|\cdot\|_E$ in place of $\|\cdot\|_W$ for the value-gradient mismatch (its Sobolev seminorm is bounded by Assumption D.1 via $\big\|\nabla_{\boldsymbol{\tau}} R^{\otimes N}\big\|_W \le L\sqrt{N}$, absorbed into the residual C_λ that contributes to the $\sigma_\zeta^{-2}(T)\cdot\mathcal{O}(1/\sqrt{N})$ tail in Eq. 7). The coefficient λ in the practical objective Eq. 8 is exactly the Young parameter; it controls the trade-off between distributional fidelity (first summand) and value alignment (second summand) and corresponds to the λ tuning knob exposed in Algorithm 1 (see Appendix H.3.1 for empirical sensitivity analysis).

Upper bound versus exact tilted-score fitting. The decomposition (50) is a strict upper bound, not an equality, so minimising $\mathcal{J}_{MF\text{-}V}^N$ minimises an upper bound on the genuine value-tilted score loss $\big\|\mathbf{s}_\theta - (\nabla \log \zeta_{T-t}^{\otimes N} + \frac{1}{\alpha}\nabla_{\boldsymbol{\tau}} R^{\otimes N})\big\|_W^2$. The bound is valid, but the minimiser of the bound is

not the minimiser of the target: minimising the two-field objective of Eq. 8 pointwise gives the shrunk tilted score $s^\star = \frac{\alpha}{\alpha+\lambda}(\nabla \log \zeta_t + \frac{\lambda}{\alpha}\nabla_\tau R)$ of Eq. 9, at effective temperature $\alpha_{\text{eff}} = \alpha/\lambda$. Two consequences are worth recording, since the second is easy to get wrong. First, the excess risk of the compromise relative to the exact target tends to $\alpha^{-2}\|\nabla_\tau R\|^2$ as $\lambda \to 0$, rather than vanishing as the bound gap $O(\lambda)$ might suggest: shrinking the fit does not remove the mismatch in the direction $\nabla_\tau R$. Second, the remedy is local and cheap: training the single squared error against $\nabla \log \zeta_t + \alpha^{-1}\nabla_\tau R$ directly makes the minimiser exactly the tilted score (MF-VSM-EXACT, Table 5), after which the upper-bound step leaves the chain entirely and ϵ^{score} in Theorem 4.2 is measured against the target the algorithm actually fits. We retain Eq. 8 as the deployed default on cost grounds, and note that because α and λ are jointly tuned, the reported optimum (α=1, λ=0.1) places α_{eff} where the App. H.3.1 sweeps locally put it—which is why the compromise is not visible in the returns alone.

Substituting the score-network decomposition (50) into the integrated Sobolev bound (47), with the $\mathcal{M}_2/\sqrt{ND_\tau}$ prefactor derived in (2.a)–(2.c), yields the bound in Eq. 7; the residual C_λ contributes the $\sigma_\zeta^{-2}(T) \cdot \mathcal{O}(1/\sqrt{N})$ tail (since $\left\|\nabla_{\boldsymbol{\tau}} R^{\otimes N}\right\|_W \le L\sqrt{N}$ by Assumption D.1 and $\sigma_\zeta^{-2}(T)$ scales the conversion from W to E norm at the smallest noise level). □

E.3 Proof of Theorem D.7: Concentration of MF-VSM

The preceding proposition relates the value-weighted objective to the target score. We next control its finite-population fluctuation; the interaction term reuses the coupling estimate from Appendix E.1.

Proof. The proof decomposes the finite-N deviation into an interaction error (controlled by the Gronwall estimate) and a statistical error (controlled by Bernstein's inequality), then shows the value-weighting terms preserve the $\tilde{\mathcal{O}}(1/\sqrt{N})$ concentration rate.

Step 1: Decomposition. Write the deviation as:

$$\mathbb{E}_t F_t^V(\boldsymbol{\tau}_t^N) - \mathcal{J}_{MF\text{-}V}(1,\theta,\mu) = \underbrace{\mathbb{E}_t[F_t^V(\boldsymbol{\tau}_t^N) - F_t^V(\bar{\boldsymbol{\tau}}_t^N)]}_{\text{(I) Interaction error}} + \underbrace{\mathbb{E}_t[F_t^V(\bar{\boldsymbol{\tau}}_t^N)] - \mathcal{J}_{MF\text{-}V}(1,\theta,\mu)}_{\text{(II) Statistical error}}, \quad (51)$$

where $\bar{\boldsymbol{\tau}}_t^N$ denotes N i.i.d. copies from the mean-field μ_t.

Step 2: Bounding Term (I). By the Lipschitz continuity of F_t^V (guaranteed by Assumptions D.1–D.2) and the Gronwall estimate in Eq. 37:

$$\mathbb{E}|F_t^V(\boldsymbol{\tau}_t^N) - F_t^V(\bar{\boldsymbol{\tau}}_t^N)| \le L_F\sqrt{\frac{C_{D_\tau}}{N}}\exp(LT). \quad (52)$$

Step 3: Bounding Term (II). Since $\bar{\boldsymbol{\tau}}_t^1, \ldots, \bar{\boldsymbol{\tau}}_t^N$ are i.i.d., $F_t^V(\bar{\boldsymbol{\tau}}_t^N) = \frac{1}{N}\sum_{i=1}^N g(\bar{\boldsymbol{\tau}}_t^i)$ for the per-agent integrand

$$g(\boldsymbol{\tau}) = \|\mathbf{s}_\theta(t,\boldsymbol{\tau}) - \nabla \log \zeta_t(\boldsymbol{\tau})\|_W^2 + \tfrac{\lambda}{\alpha}\|\mathbf{s}_\theta(t,\boldsymbol{\tau}) - \nabla_{\boldsymbol{\tau}} R(\boldsymbol{\tau})\|_E^2. \quad (53)$$

We verify that g has a sub-exponential tail under our standing assumptions. First, by Assumption D.1 (Lipschitz score network and reward), each summand in (53) is bounded by a polynomial in $\|\boldsymbol{\tau}\|$: $\|\mathbf{s}_\theta(t,\boldsymbol{\tau})\| \le L\|\boldsymbol{\tau}\| + \|\mathbf{s}_\theta(t,0)\|$ and $\|\nabla \log \zeta_t(\boldsymbol{\tau})\| \le L\|\boldsymbol{\tau}\|/\sigma_t^2$ (standard Gaussian-like reference measure), giving $g(\boldsymbol{\tau}) \le C(t)(1+\|\boldsymbol{\tau}\|^2)$ for a deterministic $C(t) < \infty$ on the inference noise interval $[t_0, T]$. Second, under the log-Sobolev Assumption D.3, the reference measure ζ_t satisfies a quadratic transportation inequality (cf. Otto–Villani (Otto & Villani, 2000, Theorem 1)), which by Herbst's argument ((Ané et al., 2000, Section 5.1)) implies that $\|\boldsymbol{\tau}\|^2$ is sub-exponentially concentrated under ζ_t:

$$\mathbb{P}_{\zeta_t}\left(\|\boldsymbol{\tau}\|^2 - \mathbb{E}\|\boldsymbol{\tau}\|^2 \ge u\right) \le \exp(-\kappa u/(2\mathcal{M}_2)) \qquad \forall\, u \ge 0.$$

Composing with the polynomial bound on g gives the sub-exponential tail $\mathbb{P}(|g - \mathbb{E}g| \ge u) \le 2\exp(-\kappa u/b_g)$ with $b_g = O(C(t)\mathcal{M}_2/\kappa)$. Applying Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10):

$$\mathbb{P}\left(|F_t^V(\bar{\boldsymbol{\tau}}_t^N) - \mathbb{E}[g(\bar{\boldsymbol{\tau}}_t^1)]| \ge \varepsilon\right) \le 2\exp\left(-\frac{N\varepsilon^2}{2(\sigma_g^2 + b_g\varepsilon/3)}\right), \quad (54)$$

where $\sigma_g^2 = \mathrm{Var}(g(\bar{\boldsymbol{\tau}}_t^1))$ is finite (since g is sub-exponential, in particular has finite variance) and b_g is the sub-exponential parameter as derived above.

Step 4: Combining. Combining Terms (I) and (II) via the triangle inequality and optimizing the split yields the concentration bound in Eq. 27. The key observation is that the value-weighting terms contribute additively to σ_g^2 and b_g but do not change the $\sqrt{N}$ concentration rate, since $R(\boldsymbol{\tau}^i)$ is bounded by Assumption D.2. □

E.4 Proof of Proposition 3.4: Value-weighted Subdivision

The single-level objective is now characterized. The subdivision proof applies its bound on successive noise intervals and accounts for the branching residual at each change of population size.

Proof. We derive the telescoping decomposition of the value-weighted chaotic entropy $\mathcal{H}_V^{\infty}(\mu_T)$ along a hierarchical schedule $\mathbb{N} = \{N_k\}_{k=0}^K$ with branching ratio $\mathfrak{b}$ and time partition $\mathbb{T} = \{t_k\}_{k=0}^K$.

Step 1: Per-level value-weighted entropy increment. Apply the Itô–Wentzell–Lions formula to the relative entropy $\mathcal{H}_V(\mu_t^{N_k})$ along the forward MF-SDE on the sub-interval $[t_k, t_{k+1}]$ (cf. Proposition 3.3, Eq. 43). The same value-weighting argument used to derive Eq. 8 from the chaotic entropy yields the per-level increment

$$\mathcal{H}_V^{N_k}(\mu_{t_{k+1}}^{N_k}) - \mathcal{H}_V^{N_k}(\mu_{t_k}^{N_k}) \;\le\; \frac{\mathcal{M}_2}{\sqrt{N_k D_\tau}} \cdot \mathcal{J}_{MF\text{-}V}\big(N_k, \theta, \nu_{[t_k,t_{k+1}]}^{N_k}\big) \;+\; \sigma_\zeta^{-2}(t_{k+1})\,\mathrm{E}(N_k), \quad (55)$$

where the first term is the value-weighted score-matching contribution at level k (under LSI, by the same argument as Proposition 3.3), and the second term $\mathrm{E}(N_k) = \mathcal{O}(1/\sqrt{N_k})$ is the residual cardinality error (entropic PoC at level k with $M = 1$, by Theorem 4.1).

Step 2: Branching cost between levels. At step t_k, the agent population grows from N_k to $N_{k+1} = \mathfrak{b}N_k$ via the branching map Ψ^θ (Eq. 12). By the data-processing inequality applied to the pushforward $(\mathbf{Id}^{\otimes(\mathfrak{b}-1)} \otimes \Psi^\theta)_\#$, and the reducibility assumption (Assumption D.5, Eq. 24, which provides the propagation-of-chaos estimate at the matched level),

$$\mathcal{H}_V^{N_{k+1}}(\mu_{t_k}^{N_{k+1}}) \;\le\; \mathcal{H}_V^{N_k}(\mu_{t_k}^{N_k}) \;+\; \mathrm{E}(N_{k+1}), \quad (56)$$

where $\mathrm{E}(N_{k+1}) = \mathcal{O}(1/\sqrt{N_{k+1}})$ absorbs the branching residual. The factor $1/(\mathfrak{b}\sqrt{N_{k+1}})^k$ in the final bound emerges from telescoping the per-level coefficients across K branching events (each branching at level $j \le k$ contributes a multiplicative factor $1/\sqrt{N_{j+1}}$ via the entropic PoC bound, while $\mathfrak{b}^{-k}$ is the geometric decay of agent budget per level).

Step 3: Telescoping. Summing (55) and (56) across $k = 0, \dots, K$ and taking the limit $K \to \infty$:

$$\mathcal{H}_V^{\infty}(\mu_T) \;\le\; \lim_{K\to\infty} \sum_{k=0}^{K} \left[\sigma_\zeta^{-2}(t_{k+1})\,\mathrm{E}(N_{k+1}) \;+\; \frac{\mathcal{M}_2}{\sqrt{D_\tau}} \left(\frac{1}{\mathfrak{b}\sqrt{N_{k+1}}} \right)^k \mathcal{J}_{MF\text{-}V}\big(N_k, \theta, \nu_{[t_k,t_{k+1}]}^{N_k}\big) \right],$$

which is exactly Eq. 10. This bound is for the *theoretically optimal* weights $(\mathfrak{b}\sqrt{N_{k+1}})^{-k}$, while the objective we minimise, $(\mathbf{P_V})$ in Eq. 11, uses $\mathfrak{b}^{-k}$. Because the two weight sequences are ordered pointwise in k and each per-level objective is non-negative, $(\mathbf{P_V})$ majorises the theoretically weighted sum (Eq. 22, Remark B.3). Minimising the implemented loss therefore controls a valid upper bound on the quantity bounded here, so Theorem 4.2 applies verbatim. The resulting majorisation gap is quantitative and measured at $1.49\times$ on trained checkpoints. □

E.5 Proof of Theorem 4.2: Hierarchical Approximation

We combine the per-level control with the finite-population and offline-shift charges to pass from generated plans to the end-to-end welfare statement.

Proof. The proof combines the performance difference lemma and Pinsker's inequality from RL theory with the entropy subdivision from Proposition 3.4 (proved in Appendix E.4), quantitative CLT bounds for the mean-field approximation, and offline RL distribution shift analysis.

Step 1: From Entropy to Policy Suboptimality. By the performance difference lemma (Kakade & Langford, 2002) applied to the joint trajectory distribution and Pinsker's inequality:

$$J(\pi^*) - J(\hat{\pi}_\theta) = \mathbb{E}_{\hat{\mu}_T}\left[\sum_{h=0}^{H-1} \gamma^h r(\mathbf{s}_h, \mathbf{a}_h, \bar{\mu}_h)\right] - \mathbb{E}_{\mu_T^*}\left[\sum_{h=0}^{H-1} \gamma^h r(\mathbf{s}_h, \mathbf{a}_h, \bar{\mu}_h)\right]$$
$$\leq \frac{2r_{\max}}{1-\gamma} D_{TV}(\hat{\mu}_T \,\|\, \mu_T^*) \;\leq\; \frac{2r_{\max}}{1-\gamma}\sqrt{\tfrac{1}{2}\mathcal{H}(\hat{\mu}_T \,\|\, \mu_T^*)}, \tag{57}$$

where the first inequality uses the standard performance-difference identity together with $|\sum_h \gamma^h r| \leq r_{\max}/(1-\gamma)$ and the symmetry of total variation $D_{TV}(P\|Q) = D_{TV}(Q\|P)$, and the second is Pinsker's inequality $D_{TV}(P\|Q) \leq \sqrt{(1/2)\mathcal{H}(P\|Q)}$ applied with $P = \hat{\mu}_T$ (the generated distribution) and $Q = \mu_T^*$ (the target). We chose this direction of KL—$\mathcal{H}(\hat{\mu}_T\|\mu_T^*)$, sometimes called the *forward* KL—because (a) it is the quantity directly controlled by the score-matching objective in Proposition 3.4 (which bounds $\mathcal{H}(\hat{\mu}_T\|\mu_T^*)$ via the sub-divided value-weighted entropy decomposition), and (b) Pinsker's inequality in this direction is sharp under the bounded-density assumption implied by Assumption D.2. The prefactor $r_{\max}/(1-\gamma)$ already absorbs the discounted reward sum—no additional H factor is needed.

Step 2: Entropy Decomposition via Girsanov and Subdivision. We bridge the entropy $\mathcal{H}(\hat{\mu}_T\|\mu_T^*)$ to a measurable score-matching quantity via Girsanov, and then refine the result through the hierarchical decomposition of Proposition 3.4. Note that Proposition 3.4 bounds the value-weighted entropy $\mathcal{H}_V^\infty(\mu_T)$, which is *not* the same object as $\mathcal{H}(\hat{\mu}_T\|\mu_T^*)$; the two are connected by the Girsanov step below rather than identified, and it is the Girsanov step that supplies the score-matching form of Terms (a)–(b).

(2.a) Global Girsanov bound (Term (a)). The reverse-time SDE driven by $\mathbf{s}_\theta$ produces $\hat{\mu}_T$, while the value-tilted target $\mu_T^* \propto \mu_T \exp(R/\alpha)$ has reverse-time score $\nabla \log \mu_t^* = \nabla \log \zeta_t + \frac{1}{\alpha}\nabla_{\boldsymbol{\tau}} R$ along the same backward dynamics (cf. Eq. 79 in the proof of Theorem 4.6). Both reverse SDEs share the same diffusion coefficient σ_t on the operational interval $[t_{\min}, T]$, so the Girsanov theorem (Karatzas & Shreve, 1991, Theorem 5.1) (with the Novikov-condition justification of Theorem 4.6, Step 2(B.i)) yields

$$\mathcal{H}(\hat{\mu}_T \,\|\, \mu_T^*) \;\leq\; \tfrac{1}{2}\int_{t_{\min}}^{T} \sigma_t^2\, \mathbb{E}\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t) - \nabla \log \mu_t^*(\boldsymbol{\tau}_t)\|^2\, dt \;\leq\; C_\sigma\, \epsilon_{\text{global}}^{score}(\theta), \tag{58}$$

where $C_\sigma = \frac{1}{2}\int_{t_{\min}}^{T} \sigma_t^2\, dt$ is the schedule constant and the second inequality uses $\sigma_t^2 \leq 2C_\sigma$ pointwise to pull the supremum out of the integral (an upper bound suffices). Combining with the Pinsker step of Step 1 directly gives Term (a) of Eq. 14.

(2.b) Hierarchical refinement (Term (b)). Proposition 3.4 (and its proof, Eq. 55) decomposes the integrated L^2 score error along the subdivision schedule:

$$C_\sigma\, \epsilon_{\text{global}}^{score}(\theta) \;\leq\; C' \sum_{k=0}^{K} \frac{(\epsilon_k^{score})^2}{\mathfrak{b}^k} \;+\; \sum_{k=0}^{K} \mathrm{E}(N_{k+1}), \tag{59}$$

where ϵ_k^{score} is the per-level score error on $[t_k, t_{k+1}]$, $\mathrm{E}(N_{k+1}) = \mathcal{O}(1/\sqrt{N_{k+1}})$ is the per-level cardinality residual, and $C' > 0$ absorbs the schedule weights σ_t^2 on each subinterval. Applying $\sqrt{\sum_k a_k} \leq \sum_k \sqrt{a_k}$ to the first sum yields the per-level form $\sum_k \mathfrak{b}^{-k/2}\epsilon_k^{score}$, which is Term (b). The cardinality residuals $\sum_k \sqrt{\mathrm{E}(N_{k+1})} = \mathcal{O}(N_0^{-1/4})$ are dominated by—and absorbed into—Term (c), since the leading mean-field contribution $H^2/\sqrt{N}$ in Term (c) controls all $1/\sqrt{N_k}$ residuals up to a multiplicative H factor that is folded into C_6.

(2.c) Why both (a) and (b) appear. Eqs. (58) and (59) provide two *complementary* upper bounds on the same score-matching contribution: (a) is the un-decomposed Girsanov bound, (b) is its hierarchical refinement under the coarse-to-fine training schedule. The total bound in Eq. 14 keeps both for interpretability—(a) is what an outside observer can measure directly via the global score-matching loss, while (b) reveals the per-level structure that the training algorithm actually exploits. Adding them is at most a factor-2 loss relative to the tighter of the two; in any quantitative instantiation either (a) or (b) alone, whichever is smaller, can replace their sum.

Step 3: Mean-field Approximation Error (from H to H^2 under Assumption D.6). The error from using $N < \infty$ agents enters the analysis through the reward and dynamics evaluated at the empirical mean field $\bar{\mu}_h^N$ versus the true mean-field μ_h. Because both the reward r and the reverse-time SDE drift are L_{eff}-Lipschitz in the mean-field argument (Assumptions D.1 and D.6), we only need to control *linear functionals of the mean field*, not full $\mathcal{W}_2$. We use the following dimension-free CLT for Lipschitz observables:

$$\mathbb{E}\Big|\langle f, \bar{\mu}_h^N\rangle - \langle f, \mu_h\rangle\Big| \;\le\; \frac{C\,\|f\|_{\text{Lip}}}{\sqrt{N}} \qquad \text{(one-step Lipschitz CLT)}, \tag{60}$$

which follows from Hoeffding's/Bernstein's inequality applied to the bounded Lipschitz observable f averaged over N exchangeable agents, and *is dimension-independent* (cf. Bolley et al. (2007, Theorem 3.1)). Crucially, this avoids the dimension-dependent rate N^{-1/d_s} of full-$\mathcal{W}_2$ Fournier–Guillin (Fournier & Guillin, 2015) for $d_s > 4$, which would otherwise degrade catastrophically on Battle ($d_s \approx 9$). The full-$\mathcal{W}_2$ formulation enters only through the trajectory-level relative-entropy bound of Theorem 4.1, which is dimension-free under LSI; the linear-functional formulation (60) suffices for all subsequent Lipschitz-induced bounds on reward and drift.

The relevant constant for trajectory-level error *compounding* is the effective Lipschitz L_{eff} of the reverse-time SDE drift (Assumption D.6): an error ε_h at step h perturbs step $h+1$ by at most $L_{\text{eff}}\varepsilon_h$, step $h+2$ by at most $L_{\text{eff}}^2\varepsilon_h$, and so on. A discrete Gronwall inequality converts (60) into

$$\mathbb{E}\Big[\max_{0\le h\le H-1}\big|\langle r, \bar{\mu}_h^N\rangle - \langle r, \mu_h\rangle\big|\Big] \;\le\; \Big(\sum_{h=0}^{H-1}(1+L_{\text{eff}})^h\Big)\cdot\frac{C\,L\,r_{\max}}{\sqrt{N}} \;\le\; e^{c_0}\,\frac{C\,H\,L\,r_{\max}}{\sqrt{N}}, \tag{61}$$

where the last inequality uses $(1+L_{\text{eff}})^h \le e^{L_{\text{eff}}h} \le e^{c_0}$ and bounds the geometric sum by $H\cdot e^{c_0}$ *by virtue of Assumption D.6*. Plugging (61) back into the finite-horizon reward accumulation yields

$$\text{Mean-field error} \;\le\; \underbrace{H}_{\substack{\text{reward sum}\\ \text{(finite-horizon)}}} \cdot \underbrace{e^{c_0}\,\frac{C\,H\,L\,r_{\max}}{\sqrt{N}}}_{\text{Gronwall-accumulated CLT}} \;=\; C_6\,e^{c_0}\,\frac{r_{\max}\,H^2\,L_{\text{eff}}}{\sqrt{N}}, \tag{62}$$

where we have absorbed $L \le L_{\text{eff}}\cdot\mathcal{O}(1)$ (the bare reward Lipschitz L is upper-bounded by the SDE-drift Lipschitz L_{eff} up to a noise-schedule-dependent constant; the gap is folded into C_6). The H^2 factor has two clearly separated origins: the *outer* H from summing finite-horizon rewards (we use the bound $\sum_{h<H}\gamma^h \le H$, which is tighter than $1/(1-\gamma)$ in the finite-horizon regime $H \le 1/(1-\gamma)$ relevant to Figure 3b), and the *inner* H from Gronwall's accumulation. **Why H^2 rather than $e^{L_{\text{eff}}H}$:** Assumption D.6 restricts $L_{\text{eff}}H \le c_0$, making $e^{L_{\text{eff}}H} \le e^{c_0} = O(1)$, so the geometric sum $\sum_{h<H}(1+L_{\text{eff}})^h$ collapses to $H\cdot O(1)$ instead of blowing up. Without this assumption, we would have $e^{LH}/\sqrt{N}$, which experiments rule out (Figure 3b fits an exponent $b \in [1.79, 2.21]$, statistically incompatible with exponential blow-up).

Reconciling the prefactor with Term (a)/(b). Terms (a) and (b) of Eq. 14 use the discounted prefactor $r_{\max}/(1-\gamma)$ obtained from $\sum_h \gamma^h \le 1/(1-\gamma)$, while Term (c) uses $r_{\max}H$ obtained from $\sum_h \gamma^h \le H$. Both bounds are valid; we use the tighter one in each context. In the finite-horizon regime $H \le 1/(1-\gamma)$ (which is the regime of all our experiments and of practical interest), $H \le 1/(1-\gamma)$, so Term (c) is no larger than what we would obtain by using $r_{\max}/(1-\gamma)$ uniformly; conversely, in the infinite-horizon limit $H \to \infty$, the second H in Term (c) is replaced by $1/(1-\gamma)$ via the same dominated convergence argument. Either reading reproduces the empirical H^2 scaling validated in Figure 3b.

Step 4: Offline Distribution Shift (Per-agent Marginal is N-free). We separate the proof of N-independence of the offline-shift term into a self-contained lemma.

Lemma E.1 (Per-agent reduction of the offline-shift charge)**.** *Let π_θ be the policy induced by sampling from the trained score network $\mathbf{s}_\theta$ in the reverse SDE driven by the empirical mean field $\bar{\mu}_h^N$. Let $\nu_\beta, \nu^* \in \mathcal{P}(\mathcal{X})$ be the per-agent trajectory marginals of Eq. 26, and $\epsilon_{offline} := D_{TV}(\nu_\beta \,\|\, \nu^*)$. Suppose the score network is parameterized so that $\mathbf{s}_\theta(t, \boldsymbol{\tau}^i; \bar{\mu}_t^N)$ depends on the joint state $\boldsymbol{\tau}^N$ only through a single-agent trajectory $\boldsymbol{\tau}^i$ and the empirical mean field $\bar{\mu}_t^N$ (permutation-equivariance; Assumption D.5). Then the joint-trajectory suboptimality contribution from offline shift, $\Delta_{\text{shift}} :=$*

$|J^{MF}(\pi^*) - J^{MF}(\pi_\theta)|_{\text{shift}}$, *satisfies*

$$\Delta_{\text{shift}} \;\le\; \frac{2r_{\max}}{1-\gamma}\,\epsilon_{offline} \;+\; \frac{r_{\max} H\, L\, e^{c_0}}{\sqrt{N}}\,\sqrt{2(C_1 + C_2 H L_{\text{eff}}^2)}, \tag{63}$$

where the $1/\sqrt{N}$ *residual is absorbed into Term (c).*

Proof of Lemma E.1. Let $\bar{\mu}_\beta^N$ (resp. $\bar{\mu}_*^N$) denote the random empirical mean-field flow generated by N agents acting under π_β (resp. π^*), with conditional single-agent laws $\pi_\beta(\boldsymbol{\tau} \mid \bar{\mu})$, $\pi_*(\boldsymbol{\tau} \mid \bar{\mu})$. The per-agent marginals of Eq. 26 satisfy $\nu_\beta(\boldsymbol{\tau}) = \mathbb{E}_{\bar{\mu}\sim\pi_\beta^N}[\pi_\beta(\boldsymbol{\tau} \mid \bar{\mu})]$ and $\nu^*(\boldsymbol{\tau}) = \mathbb{E}_{\bar{\mu}\sim\pi_*^N}[\pi_*(\boldsymbol{\tau} \mid \bar{\mu})]$, so $\epsilon_{offline} \in [0,1]$ is intrinsic to the dataset–policy pair.

(i) Cross-agent factorization of expected reward. Because the social-welfare reward is a per-agent average $J^{MF}(\pi,\mu) = (1/N)\sum_i J_i^{MF}(\pi,\mu)$ (Eq. 2) and the score network is permutation-equivariant, exchangeability of the N-agent joint laws (under either π^* or π_θ) implies

$$J^{MF}(\pi) \;=\; \mathbb{E}_{\boldsymbol{\tau}\sim\nu_\pi^{(1)}}\left[\sum_{h=0}^{H-1}\gamma^h r_h(\boldsymbol{\tau}_h)\right], \tag{64}$$

where $\nu_\pi^{(1)} \in \mathcal{P}(\mathcal{X})$ is the per-agent marginal of the joint law ν_π^N. Hence

$$|J^{MF}(\pi^*) - J^{MF}(\pi_\theta)| \;\le\; \frac{r_{\max}}{1-\gamma}\, D_{TV}\big(\nu_{\pi^*}^{(1)} \,\|\, \nu_{\pi_\theta}^{(1)}\big). \tag{65}$$

The right-hand side depends on *per-agent* marginals only.

(ii) Decomposition of the per-agent TV via offline shift and PoC residual. By the triangle inequality,

$$D_{TV}\big(\nu_{\pi^*}^{(1)} \,\|\, \nu_{\pi_\theta}^{(1)}\big) \;\le\; \underbrace{D_{TV}\big(\nu_{\pi^*}^{(1)} \,\|\, \nu^*\big)}_{\text{(I) MF-projection}} + \underbrace{D_{TV}\big(\nu^* \,\|\, \nu_\beta\big)}_{=\epsilon_{offline}} + \underbrace{D_{TV}\big(\nu_\beta \,\|\, \nu_{\pi_\theta}^{(1)}\big)}_{\text{(III) score reverse-coverage}}. \tag{66}$$

Term (I) is bounded by combining Theorem 4.1 (with $M = 1$, giving the per-agent KL bound directly without invoking Talagrand) with Pinsker's inequality $D_{TV} \le \sqrt{D_{KL}/2}$:

$$D_{TV}\big(\nu_{\pi^*}^{(1)} \,\|\, \nu^*\big) \;\le\; \sqrt{\tfrac{1}{2}\mathcal{H}\big(\nu_{\pi^*}^{(1)} \,\|\, \nu^*\big)} \;\le\; \sqrt{\frac{e^{c_0}(C_1 + C_2 H L_{\text{eff}}^2)}{2N}}.$$

Note that we use the entropic PoC bound from Theorem 4.1 *directly* (which is dimension-free under LSI), rather than going through $\mathcal{W}_2$ and Talagrand—this is essential because Talagrand under LSI gives $\mathcal{W}_2^2 \le 2\mathcal{H}/\kappa$ (not the converse), so it cannot be used to upgrade $\mathcal{W}_2$ bounds into KL bounds. Term (III) is bounded by the score-matching reverse coverage: the score network targets ν_{BR}^{soft} (the value-tilted reweighting of ν_β, Eq. 79), and the reverse SDE driven by $\mathbf{s}_\theta$ produces a per-agent marginal $\nu_{\pi_\theta}^{(1)}$ that is $\sqrt{\epsilon^{score}/2}$-close to ν_{BR}^{soft} in TV (Girsanov + Pinsker, see proof of Theorem 4.6 Step 2). Since ν_{BR}^{soft} is absolutely continuous w.r.t. ν_β with density $e^{R/\alpha}/Z$ bounded by $e^{r_{\max}H/((1-\gamma)\alpha)}$, this term is absorbed into the score-error budget of Term (a) (and not into $\epsilon_{offline}$).

Combining (I)+(II)+(III) into (65), the offline-shift contribution is exactly $\frac{r_{\max}}{1-\gamma}\cdot\epsilon_{offline}$ (with constant 2 from the conservative-direction Pinsker convention used in Step 1), plus a $1/\sqrt{N}$ residual from (I) which we charge to Term (c). This proves (63). □

Remark on tensorization and the N-independence. A naive tensorization of TV would give $D_{TV}(\nu_\beta^{\otimes N} \,\|\, \nu^{*\otimes N}) \le N\,\epsilon_{offline}$, growing linearly in N. Lemma E.1 exploits two structural properties: (i) the social-welfare reward is a per-agent *average* (Eq. 64), so cross-agent correlations cancel; (ii) the score network is *permutation-equivariant* and only consumes per-agent inputs, so its training error is intrinsically per-agent. The $1/\sqrt{N}$ residual from (I) is a vanishing finite-population correction absorbed into Term (c) (Step 3). Thus $\epsilon_{offline}$ is a per-agent dataset-coverage error independent of N, while the PoC residual carries the remaining N dependence.

Combining Steps 1–4 (with Lemma E.1 controlling Step 4) yields the bound in Eq. 14. □

E.6 Proof of Corollary 4.4: Scalability

Once the hierarchical terms are separated, the scalability statement follows by holding the score and offline terms fixed and reading off the population dependence.

Proof. This follows directly from Theorem 4.2 by noting that:

1. Term (a) is $\mathcal{O}(\epsilon^{score})$, independent of N.
2. Term (b) is $\mathcal{O}\left(\sum_k \mathfrak{b}^{-k/2}\epsilon_k^{score}\right) = \mathcal{O}(\epsilon^{score})$ for geometric decay.
3. Term (c) is $\mathcal{O}(1/\sqrt{N})$, which *decreases* with N.
4. Term (d) is $\mathcal{O}(\epsilon_{offline})$, independent of N by the PoC property.

Summing the contributions yields $J(\pi^*) - J(\hat{\pi}_\theta) = \mathcal{O}(1/\sqrt{N}) + \mathcal{O}(\epsilon^{score}) + \mathcal{O}(\epsilon_{offline})$, so the planning error is independent of N apart from the improving Term (c). □

E.7 Additional Theoretical Results

The welfare proof is complete. The next two propositions separately characterize the ideal branching map and contrast the resulting mean-field objective with a joint-space diffuser; neither is needed as an extra premise for the hierarchical bound.

E.7.1 Optimal Agent Branching Function

Proposition E.2. *(Optimal Agent Branching). The optimal agent branching function Ψ^* at step t_k minimizes the Wasserstein distance between the branched distribution and the target:*

$$\Psi^* = \arg\min_{\Psi} \mathcal{W}_2^2\left((\mathbf{Id}^{\otimes(\mathfrak{b}-1)} \otimes \Psi)_{\#}\nu_{t_k}^{N_k},\ \nu_{t_k}^{N_{k+1}}\right). \tag{67}$$

Under the exchangeability assumption, the optimal branching admits the Monge-Ampère characterization:

$$\Psi^*(\boldsymbol{\tau}^{N_k}) = \nabla\varphi^*(\boldsymbol{\tau}^{N_k}), \tag{68}$$

where φ^* solves the Monge-Ampère equation $\det\left(\nabla^2\varphi\right) = \varrho_{t_k}^{N_k}/\varrho_{t_k}^{N_{k+1}}(\nabla\varphi)$ with appropriate boundary conditions.

E.7.2 Comparison with Naive Approaches

Proposition E.3. *(Comparison with Joint Diffuser). For a standard diffuser operating on the joint space $\mathbb{R}^{ND_\tau}$, the score matching error satisfies:*

$$\mathcal{J}_{SM}^{Joint}(\theta) \geq C \cdot ND_\tau \cdot \inf_\theta \mathbb{E}[\|\mathbf{s}_\theta - \nabla\log\zeta_t\|^2], \tag{69}$$

which grows linearly in N. In contrast, the MF-VSM objective satisfies:

$$\mathcal{J}_{MF\text{-}V}^N(\theta) \leq \frac{C'\sqrt{D_\tau}}{\sqrt{N}} \cdot \inf_\theta \mathbb{E}[\|\mathbf{s}_\theta - \nabla\log\zeta_t\|_W^2] + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right), \tag{70}$$

which decreases with N, demonstrating the fundamental advantage of the mean-field approach.

E.8 Proofs for Game-theoretic Analysis (Appendix D.4)

We present complete proofs of the game-theoretic results. The exploitability proof reduces an N-player deviation to a mean-field one and uses the consistency estimate proved immediately afterward. Monotonicity then combines with consistency to establish convergence and its ε-Nash corollary; the final proposition separately bounds the welfare–Nash efficiency gap. The proofs adapt the N-player to MFG convergence framework of Lacker (2016) and Fischer (2017), the monotone operator theory of Lasry & Lions (2007), and the master equation approach of Cardaliaguet et al. (2019) to the discrete-time trajectory diffusion setting. Each adaptation requires handling the temporal coupling and the value-weighted score matching structure specific to MF-Diffuser; the conceptual MFG-vs-MFRL distinction itself is in Appendix D.4 (Remark on MFG vs. MFRL).

E.8.1 Proof of Theorem 4.6: Exploitability Bound

Proof. The proof decomposes the N-player exploitability into a mean-field exploitability and an N-to-MFG approximation error, then bounds each term using the PoC results (Theorem 4.1) and the properties of the MF-VSM objective.

Step 1: N-player to Mean-field Reduction.

For any deviating policy π' for agent 1 while agents $2, \dots, N$ use $\hat{\pi}_\theta$:

$$J_1(\pi', \hat{\pi}_\theta^{\otimes(N-1)}) = J^{MF}(\pi', \Phi(\hat{\pi}_\theta)) + \Delta_N(\pi'), \tag{71}$$

where $\Delta_N(\pi')$ captures the deviation from the mean-field approximation. We bound Δ_N by analyzing the effect of a single agent's deviation on the empirical measure.

Under $(\pi', \hat{\pi}_\theta^{\otimes(N-1)})$, the empirical distribution at step h is $\bar{\mu}_h^{N,\pi'} = \frac{1}{N}\delta_{\mathbf{s}_h^{1,\pi'}} + \frac{N-1}{N}\bar{\mu}_h^{(N-1),\hat{\pi}_\theta}$. Agent 1's deviation replaces one Dirac mass out of N, so by the triangle inequality and the quantitative CLT for the remaining $N-1$ exchangeable agents (following (Bolley et al., 2007)):

$$\mathbb{E}\left[\mathcal{W}_2\left(\bar{\mu}_h^{N,\pi'}, \Phi(\hat{\pi}_\theta)_h\right)\right] \le \frac{C_{d_s}}{\sqrt{N-1}} + \frac{1}{N}\mathbb{E}\left[\left\|\mathbf{s}_h^{1,\pi'} - \mathbf{s}_h^{1,\hat{\pi}_\theta}\right\|\right]. \tag{72}$$

The first term is the standard CLT rate for the empirical measure of the $N-1$ conforming agents; the second is the $O(1/N)$ contribution of agent 1's deviation. By the Lipschitz continuity of the reward (Assumption D.1), the deviation from the mean-field objective accumulated over the horizon satisfies:

$$|\Delta_N(\pi')| \le \sum_{h=0}^{H-1} \gamma^h L \cdot \mathbb{E}\left[\mathcal{W}_2\left(\bar{\mu}_h^{N,\pi'}, \Phi(\hat{\pi}_\theta)_h\right)\right]. \tag{73}$$

The dynamics propagate state errors with the effective Lipschitz constant L_{eff} (Assumption D.6); applying a discrete Gronwall inequality (the same mechanism as in the proof of Theorem 4.1, Step 2) and invoking $L_{\text{eff}} H \le c_0$ to bound $(1 + L_{\text{eff}})^H \le e^{c_0}$:

$$|\Delta_N(\pi')| \le C_\Delta\, e^{c_0} \cdot \frac{r_{\max} H^2 L_{\text{eff}}}{\sqrt{N}}, \tag{74}$$

where $C_\Delta > 0$ depends on d_s and γ. The H^2 factor arises from (*outer*) the finite-horizon reward sum and (*inner*) the Gronwall accumulation, with the multiplicative e^{c_0} bounded by Assumption D.6. This adapts the N-player to MFG convergence of Lacker (2016, Theorem 2.4) from continuous-time Itô diffusions to discrete-time MDPs; the key modification is replacing the continuous Gronwall lemma with its discrete counterpart while preserving the $O(1/\sqrt{N})$ rate through the exchangeability structure (Assumption D.4).

Taking the supremum over π' in Eqs. 71–74:

$$\text{Exploit}_N(\hat{\pi}_\theta) \le \text{Exploit}^{MF}(\hat{\pi}_\theta) + \frac{2C_\Delta\, e^{c_0}\, r_{\max} H^2 L_{\text{eff}}}{\sqrt{N}}, \tag{75}$$

where the mean-field exploitability is:

$$\text{Exploit}^{MF}(\hat{\pi}_\theta) := \sup_{\pi'} \left[J^{MF}(\pi', \Phi(\hat{\pi}_\theta)) - J^{MF}(\hat{\pi}_\theta, \Phi(\hat{\pi}_\theta))\right]. \tag{76}$$

Step 2: Bounding Mean-field Exploitability via Value-weighted Score Matching.

We decompose the mean-field exploitability into three sub-terms:

$$\begin{aligned}\text{Exploit}^{MF}(\hat{\pi}_\theta) = &\underbrace{J^{MF}(\text{BR}(\Phi(\hat{\pi}_\theta)), \Phi(\hat{\pi}_\theta)) - J^{MF}(\pi_{BR}^{soft}, \Phi(\hat{\pi}_\theta))}_{\text{(A) softmax bias}} \\ &+ \underbrace{J^{MF}(\pi_{BR}^{soft}, \Phi(\hat{\pi}_\theta)) - J^{MF}(\hat{\pi}_\theta, \hat{\mu}_\theta)}_{\text{(B) generation error}} \\ &+ \underbrace{J^{MF}(\hat{\pi}_\theta, \hat{\mu}_\theta) - J^{MF}(\hat{\pi}_\theta, \Phi(\hat{\pi}_\theta))}_{\text{(C) consistency gap}},\end{aligned} \tag{77}$$

where $\pi_{BR}^{soft}(\cdot|\mathbf{s},\mu) \propto \pi_\beta(\cdot|\mathbf{s},\mu)\exp\big(Q^{MF}(\mathbf{s},\cdot,\mu)/\alpha\big)$ is the soft (entropy-regularized) best response under the RL-as-inference framework (Levine, 2018), and $\hat{\mu}_\theta$ is the mean-field distribution from the diffusion model's generated trajectories.

Bounding (A): Softmax Bias. The soft best response maximizes the entropy-regularized objective $J^{MF}(\pi,\mu) + \alpha\mathbb{E}[\mathcal{H}(\pi(\cdot|\mathbf{s},\mu))]$, where $\mathcal{H}$ denotes Shannon entropy. By the duality between hard and soft optimization (see *e.g.*, Levine (2018, Section 3)):

$$J^{MF}(\mathrm{BR}(\mu),\mu) - J^{MF}(\pi_{BR}^{soft},\mu) \le \alpha\cdot\mathbb{E}\left[\sum_{h=0}^{H-1}\frac{\gamma^h}{1-\gamma}\mathcal{H}(\pi_{BR}^{soft}(\cdot|\mathbf{s}_h,\mu_h))\right] \le \frac{\alpha\bar{\mathcal{H}}_\pi}{1-\gamma}, \tag{78}$$

where $\bar{\mathcal{H}}_\pi = \sup_{\mathbf{s},\mu}\mathcal{H}(\pi_{BR}^{soft}(\cdot|\mathbf{s},\mu))$ is the maximum policy entropy, which is finite under compactness of $\mathcal{A}$ (Assumption D.1).

Bounding (B): Generation Error via Girsanov. The MF-VSM objective (Eq. 8) trains the score network to generate trajectories from the value-weighted (*tilted*) distribution

$$\nu_{BR}^{\mathrm{soft}}(\boldsymbol{\tau}) \;\propto\; \nu_\beta(\boldsymbol{\tau})\exp\big(R(\boldsymbol{\tau})/\alpha\big), \tag{79}$$

whose induced policy is precisely π_{BR}^{soft} under the training-time mean-field. The score of ν_{BR}^{soft} is exactly the value-tilted score $\nabla\log\zeta_t + \frac{1}{\alpha}\nabla_{\boldsymbol{\tau}}R$ that the MF-VSM objective targets (modulo the Sobolev-Young decomposition of Eq. 50); we denote this target $\mathbf{s}_t^*(\boldsymbol{\tau}) \coloneqq \nabla\log\zeta_t(\boldsymbol{\tau}) + \frac{1}{\alpha}\nabla_{\boldsymbol{\tau}}R(\boldsymbol{\tau})$ and define

$$\epsilon_{SM}(\theta) \;\coloneqq\; \mathbb{E}_{t\sim p(t),\boldsymbol{\tau}_t\sim\hat{\nu}_\theta}\big[\|\mathbf{s}_\theta(t,\boldsymbol{\tau}_t) - \mathbf{s}_t^*(\boldsymbol{\tau}_t)\|^2\big]. \tag{80}$$

We work with three trajectory measures on $\mathcal{X}$ to keep the offline/online distinction explicit:

- ν_β: the offline data distribution (per-agent trajectory marginal under behavior policy π_β, cf. Lemma E.1).
- $\nu_{BR}^{\mathrm{soft}} \propto \nu_\beta\exp(R/\alpha)$: the value-tilted reweighting of *the offline base* (Eq. 79). *This is the actual training target* of the MF-VSM objective.
- $\nu_{BR}^{\mathrm{soft},*} \propto \nu^*\exp(R/\alpha)$, where ν^* is the population-level per-agent trajectory marginal under the true MFE flow. *This is the ideal target* that defines π_{BR}^{soft} on the right-hand side of (81).

Let $\hat{\nu}_\theta$ denote the law of the reverse SDE driven by $\mathbf{s}_\theta$, which approximates ν_{BR}^{soft} (the training target).

By the performance difference lemma (Kakade & Langford, 2002) *in TV form*:

$$J^{MF}(\pi_{BR}^{\mathrm{soft}},\hat{\mu}_\theta) - J^{MF}(\hat{\pi}_\theta,\hat{\mu}_\theta) \le \frac{2r_{\max}}{1-\gamma}\,D_{TV}\big(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\mathrm{soft},*}\big), \tag{81}$$

where the prefactor $r_{\max}/(1-\gamma)$ correctly captures the discounted reward range (no extraneous H). We bound this TV distance via the triangle inequality (which holds for TV but *not* for KL):

$$D_{TV}\big(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\mathrm{soft},*}\big) \;\le\; \underbrace{D_{TV}\big(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\mathrm{soft}}\big)}_{\text{(B.i) generation gap}} + \underbrace{D_{TV}\big(\nu_{BR}^{\mathrm{soft}} \,\|\, \nu_{BR}^{\mathrm{soft},*}\big)}_{\text{(B.ii) offline-shift gap}}. \tag{82}$$

The triangle inequality is used in total variation rather than in KL, because no triangle inequality holds for relative entropy in general; all subsequent KL bounds enter through Pinsker's inequality $D_{TV} \le \sqrt{D_{KL}/2}$.

(B.i) Girsanov bound on the generation gap. For two reverse-time SDEs with the *same* diffusion coefficient σ_t but drifts $b_t^\theta = f_t - \sigma_t^2\mathbf{s}_\theta(t,\cdot)$ and $b_t^* = f_t - \sigma_t^2\mathbf{s}_t^*(\cdot)$ (where $\mathbf{s}_t^*$ is the score of ν_{BR}^{soft}), the Girsanov theorem (Karatzas & Shreve, 1991, Theorem 5.1) on the *operational* interval $[t_{\min},T]$ yields

$$D_{KL}(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\mathrm{soft}}) \;\le\; \frac{1}{2}\int_{t_{\min}}^{T}\sigma_t^2\,\mathbb{E}_{\hat{\nu}_\theta}\big[\|\mathbf{s}_\theta(t,\boldsymbol{\tau}_t) - \mathbf{s}_t^*(\boldsymbol{\tau}_t)\|^2\big]\,dt \;\le\; C_\sigma(t_{\min})\,\epsilon_{SM}(\theta), \tag{83}$$

where $C_\sigma(t_{\min}) = \frac{1}{2}\int_{t_{\min}}^{T}\sigma_t^2\,dt$ is finite and bounded uniformly in $t_{\min} > 0$ for any standard noise schedule (e.g. VP/VE), and the integral is restricted to $[t_{\min},T]$ because the reverse

SDE is integrated only on this interval at inference time (we never sample at $t < t_{\min}$, so the small-noise singularity $\sigma_t^{-2} \to \infty$ as $t \to 0$ does not enter the bound). The Novikov condition $\mathbb{E}\exp\big(\frac{1}{2}\int_{t_{\min}}^{T} \sigma_t^2 \|\mathbf{s}_\theta - \mathbf{s}_t^*\|^2 dt\big) < \infty$ holds because (i) σ_t^2 is bounded on the closed interval $[t_{\min}, T]$, and (ii) the score outputs are bounded over the same interval by Assumption D.6. Combining with Pinsker:

$$D_{TV}\big(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\text{soft}}\big) \;\le\; \sqrt{\tfrac{1}{2} D_{KL}(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\text{soft}})} \;\le\; \sqrt{\tfrac{1}{2} C_\sigma(t_{\min})\, \epsilon_{SM}(\theta)}. \tag{84}$$

(B.ii) Offline-shift correction. The two tilted distributions share the exponent $\exp(R/\alpha)$ but differ in their base measure (ν_β vs. ν^*). Their Radon–Nikodym derivative is

$$\frac{d\nu_{BR}^{\text{soft}}}{d\nu_{BR}^{\text{soft},\,*}}(\boldsymbol{\tau}) \;=\; \frac{Z^{\text{soft},\,*}}{Z_{BR}^{\text{soft}}} \cdot \frac{d\nu_\beta}{d\nu^*}(\boldsymbol{\tau}). \tag{85}$$

Under bounded reward (Assumption D.2), $|R(\boldsymbol{\tau})| \le r_{\max} H/(1-\gamma)$, so the normalizer ratio is bounded by $\exp(r_{\max} H/((1-\gamma)\alpha))$. By the elementary inequality $|p - q| \le \frac{1}{2}\|p/q - 1\|_\infty \cdot \|q\|_1$ for absolutely continuous measures and the standard offline-RL pessimism inequality (Jin et al., 2021, Lemma 3.1),

$$D_{TV}\big(\nu_{BR}^{\text{soft}} \,\|\, \nu_{BR}^{\text{soft},\,*}\big) \;\le\; \exp\!\left(\frac{r_{\max} H}{(1-\gamma)\alpha}\right) \epsilon_{offline} \;=:\; C_{off}(H,\alpha)\, \epsilon_{offline}, \tag{86}$$

where we make the constant's dependence on H and α *explicit*: C_{off} is *not* a universal constant but grows exponentially in the value-tilt strength $r_{\max} H/((1-\gamma)\alpha)$. In practice this is harmless because $\alpha \sim 1$ and bounded rewards (Assumption D.2) cap the per-trajectory tilt at $e^{r_{\max} H/((1-\gamma)\alpha)}$, but the (H, α)-dependence should be tracked as a sample-complexity cost of strong tilting.

Combining (84) and (86) into (82):

$$D_{TV}\big(\hat{\nu}_\theta \,\|\, \nu_{BR}^{\text{soft},\,*}\big) \;\le\; \sqrt{\tfrac{1}{2} C_\sigma(t_{\min})\, \epsilon_{SM}(\theta)} \;+\; C_{off}(H,\alpha)\, \epsilon_{offline}. \tag{87}$$

This is the version of TV that controls the value gap (B) via (81) since π_{BR}^{soft} is defined relative to the population (online) mean field.

Bounding (C): Consistency Gap. By the L-Lipschitz continuity of r in the mean-field argument and the Gronwall-bounded $\mathcal{W}_2$ deviation under Assumption D.6:

$$|J^{MF}(\hat{\pi}_\theta, \hat{\mu}_\theta) - J^{MF}(\hat{\pi}_\theta, \Phi(\hat{\pi}_\theta))| \le \frac{r_{\max} L}{1-\gamma}\, e^{c_0} \max_h \mathcal{W}_2(\hat{\mu}_{h,\theta}, \Phi(\hat{\pi}_\theta)_h). \tag{88}$$

By Theorem D.12, the Wasserstein gap is $\mathcal{O}(\sqrt{\epsilon_{SM}} + \epsilon_{offline} + H/\sqrt{N})$, which is absorbed into the leading-order terms in (A), (B), and (75).

Step 3: Assembly.

Combining Eq. 75 with the bounds on (A) (Eq. 78), (B) (Eqs. 81, 87, in TV form), and (C) (Eq. 88):

$$\begin{aligned}\text{Exploit}_N(\hat{\pi}_\theta) \le\; & \frac{2 r_{\max}}{1-\gamma}\left[\sqrt{\tfrac{1}{2} C_\sigma(t_{\min})\, \epsilon_{SM}(\theta)} + C_{off}(H,\alpha)\, \epsilon_{offline}\right] + \frac{\alpha \bar{\mathcal{H}}_\pi}{1-\gamma} \\ & + \frac{2 C_\Delta\, e^{c_0}\, r_{\max} H^2 L_{\text{eff}}}{\sqrt{N}} + \frac{r_{\max} L\, e^{c_0}}{1-\gamma}\left(C_{10}\frac{\sqrt{\epsilon_{SM}}}{\kappa} + C_{11}\cdot \epsilon_{offline}\right).\end{aligned} \tag{89}$$

The TV form of (B), $\sqrt{\frac{1}{2} C_\sigma \epsilon_{SM}} + C_{off}\epsilon_{offline}$, is tighter than the expression $\sqrt{2(C_\sigma \epsilon_{SM} + C_{off}\epsilon_{offline})}$ that a direct KL route would give, and it is available precisely because the triangle inequality is applied in TV. Grouping yields the bound in Eq. 16 with $C_8 = 2C_\Delta$ and C_9 absorbing the constants from the offline, consistency, and RN-correction terms; the explicit (H, α)-dependence of C_{off} should be folded into C_9 in any quantitative instantiation of the bound. □

E.8.2 Proof of Theorem D.12: Mean-field Consistency

Proof. The proof establishes self-consistency by showing that the diffusion model's generated trajectories collectively induce a population distribution close to the mean-field under which they were trained. The argument uses the Girsanov theorem for measure change under SDE perturbation and the Lipschitz stability of the mean-field flow.

Step 1: Trajectory-to-Distribution Map.

Let $\hat{\nu}_\theta$ be the trajectory distribution generated by the reverse MF-SDE. The generated mean-field at step h is:

$$\hat{\mu}_{h,\theta} = \mathrm{proj}_{\mathbf{s}_h,\#}\hat{\nu}_\theta^{(1)}, \tag{90}$$

where $\hat{\nu}_\theta^{(1)}$ is the single-agent marginal and $\mathrm{proj}_{\mathbf{s}_h}$ extracts the state component at time h. The true mean-field under $\hat{\pi}_\theta$ is defined recursively by the consistency operator Φ:

$$\Phi(\hat{\pi}_\theta)_0 = \mu_0, \quad \Phi(\hat{\pi}_\theta)_{h+1} = P_\#(\hat{\pi}_\theta(\cdot|\cdot, \Phi(\hat{\pi}_\theta)_h) \otimes \Phi(\hat{\pi}_\theta)_h), \tag{91}$$

where $P_\#$ denotes the pushforward through the transition kernel.

Step 2: Wasserstein Control via Score Matching.

The key observation: if the score matching error is zero, the generated trajectories exactly sample from the offline data distribution ν_β, for which self-consistency holds by construction (the offline data is generated by agents interacting under π_β, so $\mu_{h,\beta} = \Phi(\pi_\beta)_h$). The score matching error introduces a distributional perturbation.

By the stability of SDEs under drift perturbation (Girsanov theorem, see *e.g.*, Carmona & Delarue (2018, Theorem 6.4.2)):

$$\mathcal{W}_2^2(\hat{\nu}_\theta^{(1)}, \nu_\beta^{(1)}) \le C_{Gir} \int_0^T \sigma_t^2 \, \mathbb{E}\left[\|\mathbf{s}_\theta(t, \boldsymbol{\tau}_t) - \nabla \log \zeta_t(\boldsymbol{\tau}_t)\|^2\right] dt \le C_{Gir}\bar{\sigma}^2 \, \epsilon_{SM}(\theta), \tag{92}$$

where $\bar{\sigma}^2 = \int_0^T \sigma_t^2 \, dt$ and C_{Gir} depends on the diffusion schedule.

Step 3: Propagation to Per-step Mean-field.

The projection $\mathrm{proj}_{\mathbf{s}_h}$ is 1-Lipschitz as a map on trajectory space. Combined with the Lipschitz dynamics (Assumption D.1), errors propagate through the transition kernel:

$$\mathcal{W}_2(\hat{\mu}_{h,\theta}, \mu_{h,\beta}) \le (1+L)^h \, \mathcal{W}_2(\hat{\nu}_\theta^{(1)}, \nu_\beta^{(1)}) \le (1+L)^h C_{Gir}^{1/2}\bar{\sigma}\sqrt{\epsilon_{SM}(\theta)}. \tag{93}$$

Meanwhile, the Lipschitz property of the mean-field flow operator gives:

$$\mathcal{W}_2(\Phi(\hat{\pi}_\theta)_h, \Phi(\pi_\beta)_h) \le (1+L)^h L \cdot d_\Pi(\hat{\pi}_\theta, \pi_\beta). \tag{94}$$

The policy distance is controlled by the trajectory distribution distance plus the value-weighting tilt. Since $\hat{\pi}_\theta$ is induced by the value-weighted distribution $p^*(\boldsymbol{\tau}) \propto p_\beta(\boldsymbol{\tau}) \exp(R(\boldsymbol{\tau})/\alpha)$:

$$d_\Pi(\hat{\pi}_\theta, \pi_\beta) \le C_{KL}\sqrt{\epsilon_{SM}(\theta)} + C_V \epsilon_{offline}, \tag{95}$$

where C_V accounts for the value-weighting shift (bounded by Assumption D.2).

Step 4: Triangle Inequality and Gronwall Sum.

Since $\mu_{h,\beta} = \Phi(\pi_\beta)_h$ (self-consistency of the behavior policy), the triangle inequality gives, for each h,

$$\mathcal{W}_2(\hat{\mu}_{h,\theta}, \Phi(\hat{\pi}_\theta)_h) \le \mathcal{W}_2(\hat{\mu}_{h,\theta}, \mu_{h,\beta}) + \mathcal{W}_2(\Phi(\pi_\beta)_h, \Phi(\hat{\pi}_\theta)_h), \tag{96}$$

where the first summand is bounded by $C_{Gir}^{1/2}\bar{\sigma}\sqrt{\epsilon_{SM}}$ (Girsanov, Step 2) and the second by $L(C_{KL}\sqrt{\epsilon_{SM}} + C_V\epsilon_{offline})$ (flow-Lipschitz applied to the policy distance, Step 3). Each *single-step* error is therefore bounded by $\delta := \bar{C}\sqrt{\epsilon_{SM}} + \bar{C}'\epsilon_{offline}$ (with $\bar{C}, \bar{C}'$ absorbing $C_{Gir}^{1/2}\bar{\sigma}, LC_{KL}, LC_V$).

Discrete Gronwall accumulation. The error $e_h := \mathcal{W}_2(\hat{\mu}_{h,\theta}, \Phi(\hat{\pi}_\theta)_h)$ accumulates over horizon steps because the SDE drift in the reverse process is L_{eff}-Lipschitz (Assumption D.6): an error e_{h-1} at

step $h-1$ contributes at most $L_{\text{eff}}\, e_{h-1}$ to step h, plus a fresh per-step error δ. The discrete Gronwall recursion

$$e_h \;\le\; \delta + (1 + L_{\text{eff}})\, e_{h-1}, \qquad e_0 = \delta,$$

unfolds to

$$e_h \;\le\; \delta \sum_{j=0}^{h} (1 + L_{\text{eff}})^j \;=\; \delta \cdot \frac{(1 + L_{\text{eff}})^{h+1} - 1}{L_{\text{eff}}}. \tag{97}$$

Linear-H collapse under Assumption D.6. Since $L_{\text{eff}}\, H \le c_0 = O(1)$, we have $(1 + L_{\text{eff}})^{h+1} \le e^{L_{\text{eff}}(h+1)} \le e^{c_0}$, and the geometric sum collapses to

$$\sum_{j=0}^{h} (1 + L_{\text{eff}})^j \;\le\; (h+1)\, e^{c_0} \;\le\; H\, e^{c_0}.$$

This is the source of the *linear* H factor: the geometric sum $\sum_{j<H}(1 + L_{\text{eff}})^j$ is dominated by $H \cdot \max_j (1 + L_{\text{eff}})^j = H \cdot e^{c_0}$ rather than the much larger $((1 + L_{\text{eff}})^H - 1)/L_{\text{eff}}$ which, without Assumption D.6, could be exponential in H. Plugging back into (97) and taking the $\max_h$:

$$\max_h e_h \;\le\; H\, e^{c_0}\, \delta \;=\; H\, e^{c_0} \big[\bar{C} \sqrt{\epsilon_{SM}(\theta)} + \bar{C}'\, \epsilon_{offline} \big]. \tag{98}$$

The log-Sobolev inequality (Assumption D.3) provides additional refinement on the Girsanov-side single-step error: $\mathcal{W}_2^2(\hat{\nu}_\theta^{(1)}, \nu_\beta^{(1)}) \le (1/\kappa) D_{KL}(\hat{\nu}_\theta^{(1)} \| \nu_\beta^{(1)})$, dividing $\bar{C}$ by $\sqrt{\kappa}$ (and we absorb e^{c_0} into the universal constants C_{10}, C_{11}):

$$\max_h \mathcal{W}_2(\hat{\mu}_{h,\theta}, \Phi(\hat{\pi}_\theta)_h) \;\le\; C_{10}\, \frac{H\, L\, \sqrt{\epsilon_{SM}(\theta)}}{\kappa} + C_{11}\, H\, L\, \epsilon_{offline}. \tag{99}$$

This is exactly Eq. 32; the linear H factor on the right-hand side is the Gronwall-sum collapse, *not* an exponential. □

E.8.3 Proof of Theorem D.13: Convergence under Lasry–Lions Monotonicity

Proof. The proof leverages the Lasry–Lions monotonicity condition to establish that the MF-NE fixed-point map is a contraction, then propagates approximation errors using the Banach fixed point theorem. The contraction argument follows the classical approach of Lasry & Lions (2007), adapted from continuous-time PDEs to discrete-time MDPs through the trajectory-level framework; we additionally handle the offline distribution shift and score matching approximation specific to diffusion-based planning. The adaptation from the continuous-time PDE framework of Cardaliaguet et al. (2019) to discrete-time MDPs replaces the Hamilton–Jacobi–Bellman/Fokker–Planck system with Bellman equations coupled to the mean-field flow, but the core monotonicity mechanism is preserved.

Step 1: Uniqueness of MF-NE via Banach Fixed Point.

A mean-field Nash equilibrium satisfies the fixed-point equation $\pi^{MFE} = \Gamma(\pi^{MFE})$ where $\Gamma = \text{BR} \circ \Phi$. We show Γ is a contraction under the monotonicity condition.

Consider two policies π_1, π_2 with induced mean-fields $\mu_1 = \Phi(\pi_1)$, $\mu_2 = \Phi(\pi_2)$. Define the corresponding best responses $\hat{\pi}_k = \text{BR}(\mu_k)$ for $k = 1, 2$. By the optimality of $\hat{\pi}_k$ for mean-field μ_k:

$$J^{MF}(\hat{\pi}_1, \mu_1) \ge J^{MF}(\hat{\pi}_2, \mu_1), \tag{100}$$

$$J^{MF}(\hat{\pi}_2, \mu_2) \ge J^{MF}(\hat{\pi}_1, \mu_2). \tag{101}$$

Adding Eqs. 100–101:

$$\big[J^{MF}(\hat{\pi}_1, \mu_1) - J^{MF}(\hat{\pi}_1, \mu_2)\big] + \big[J^{MF}(\hat{\pi}_2, \mu_2) - J^{MF}(\hat{\pi}_2, \mu_1)\big] \ge 0. \tag{102}$$

Expanding via the reward structure and applying the Lasry–Lions monotonicity (Assumption D.11):

$$\begin{aligned} J^{MF}(\pi, \mu_1) - J^{MF}(\pi, \mu_2) = &\sum_{h=0}^{H-1} \gamma^h \int_{\mathcal{S}} \big[r(\mathbf{s}, \pi(\mathbf{s}, \mu_1), \mu_{1,h}) - r(\mathbf{s}, \pi(\mathbf{s}, \mu_2), \mu_{2,h})\big]\, d\mu_{1,h}(\mathbf{s}) \\ &+ \sum_{h=0}^{H-1} \gamma^h \int_{\mathcal{S}} r(\mathbf{s}, \pi(\mathbf{s}, \mu_2), \mu_{2,h})\, d(\mu_{1,h} - \mu_{2,h})(\mathbf{s}). \end{aligned} \tag{103}$$

The second sum, by the monotonicity condition (Eq. 31), satisfies:

$$\sum_{h=0}^{H-1} \gamma^h \int_{\mathcal{S}} r(\mathbf{s}, \mathbf{a}, \mu_{2,h})\, d(\mu_{1,h} - \mu_{2,h})(\mathbf{s}) \leq -\lambda_{LL} \sum_{h=0}^{H-1} \gamma^h \mathcal{W}_2^2(\mu_{1,h}, \mu_{2,h}). \tag{104}$$

The first sum (the "policy-shift" contribution) is bounded via Lipschitzness of r in (a, μ) (Assumption D.1, constant L) and Lipschitzness of the soft-BR policy in the mean-field argument (constant $L_\pi \leq L$). Defining

$$\Delta_h(\mathbf{s}) \coloneqq r(\mathbf{s}, \pi(\mathbf{s}, \mu_1), \mu_{1,h}) - r(\mathbf{s}, \pi(\mathbf{s}, \mu_2), \mu_{2,h}),$$

we obtain the *linear* bound

$$\left| \int_{\mathcal{S}} \Delta_h(\mathbf{s})\, d\mu_{1,h} \right| \;\leq\; (L\, L_\pi + L)\, \mathcal{W}_2(\mu_{1,h}, \mu_{2,h}) \;\leq\; 2L^2\, \mathcal{W}_2(\mu_{1,h}, \mu_{2,h}).$$

which is *linear* in $\mathcal{W}_2$: the two Lipschitz constants enter additively through the policy shift rather than multiplicatively, so the policy-shift contribution is $L(L_\pi + 1)\mathcal{W}_2$ with $L_\pi \leq L$, not $L^2\mathcal{W}_2^2$. Multiplying by γ^h and summing, $\sum_h \gamma^h \leq 1/(1-\gamma)$ gives $\sum_h \gamma^h \cdot 2L^2\mathcal{W}_2(\mu_{1,h}, \mu_{2,h}) \leq \frac{2L^2}{1-\gamma} \max_h \mathcal{W}_2(\mu_{1,h}, \mu_{2,h})$.

The second sum, by Lasry–Lions monotonicity (Assumption D.11, Eq. 31), is bounded *quadratically* as $\leq -\lambda_{LL} \sum_h \gamma^h \mathcal{W}_2^2(\mu_{1,h}, \mu_{2,h})$.

Closure of the contraction argument. The naive substitution into (102) mixes a linear $\mathcal{W}_2$ bound (policy-shift) with a quadratic $\mathcal{W}_2^2$ bound (monotonicity), so a direct $\mathcal{W}_2$ contraction does not follow from elementary algebra. The correct closure, due to Lasry & Lions (2007) and Cardaliaguet et al. (2019), applies the monotonicity inequality to the *gradient* (force) field rather than the value, yielding a *quadratic* bound on *both* sides of the variational inequality and hence the contraction

$$\max_h \mathcal{W}_2(\Phi(\hat{\pi}_1)_h, \Phi(\hat{\pi}_2)_h) \;\leq\; \underbrace{\frac{L^2}{(1-\gamma)\,\lambda_{LL}}}_{=L_{BR}} \max_h \mathcal{W}_2(\mu_{1,h}, \mu_{2,h}), \tag{105}$$

in agreement with Eq. 33. The detailed proof (Cardaliaguet 2019, Theorem 3.7 in the discrete-time setting) tracks the bilinear form $\langle \nabla J(\pi_1, \mu_1) - \nabla J(\pi_2, \mu_2), \pi_1 - \pi_2 \rangle$ and uses strict monotonicity to extract $\lambda_{LL}\mathcal{W}_2^2$ on the LHS and $L^2\mathcal{W}_2^2$ on the RHS, with the $1/(1-\gamma)$ factor from the discounted reward sum. Therefore $\Gamma = \mathrm{BR} \circ \Phi$ is a contraction with coefficient $L_{BR} = L^2/((1-\gamma)\lambda_{LL})$ on the complete metric space $(\mathcal{P}_2(\mathcal{S})^H, \max_h \mathcal{W}_2)$. When $\lambda_{LL}\,(1-\gamma) > L^2$ (*i.e.*, , $L_{BR} < 1$), the Banach fixed-point theorem guarantees existence and uniqueness of the MF-NE.

Step 2: Distance to MF-NE.

Define the *Nash residual*: $\eta(\hat{\pi}_\theta) = d_\Pi(\hat{\pi}_\theta, \Gamma(\hat{\pi}_\theta))$, measuring how far $\hat{\pi}_\theta$ is from being a fixed point. By the contraction:

$$\begin{aligned} d_\Pi(\hat{\pi}_\theta, \pi^{MFE}) &\leq d_\Pi(\hat{\pi}_\theta, \Gamma(\hat{\pi}_\theta)) + d_\Pi(\Gamma(\hat{\pi}_\theta), \Gamma(\pi^{MFE})) \\ &\leq \eta(\hat{\pi}_\theta) + L_{BR} \cdot d_\Pi(\hat{\pi}_\theta, \pi^{MFE}). \end{aligned} \tag{106}$$

Rearranging:

$$d_\Pi(\hat{\pi}_\theta, \pi^{MFE}) \leq \frac{\eta(\hat{\pi}_\theta)}{1 - L_{BR}}. \tag{107}$$

The Nash residual is bounded by combining the softmax bias and generation quality from the proof of Theorem 4.6 (Step 2):

$$\eta(\hat{\pi}_\theta) \leq C' \sqrt{\epsilon_{SM}(\theta)} + C'' \epsilon_{offline} + C_\alpha \sqrt{\frac{\alpha \bar{\mathcal{H}}_\pi}{1-\gamma}}. \tag{108}$$

Propagating through the Lipschitz map Φ and absorbing the α-dependent term (which can be made arbitrarily small):

$$\max_h \mathcal{W}_2(\hat{\mu}_{h,\theta}, \mu_h^{MFE}) \leq \frac{L}{1 - L_{BR}} \left[C_{12} \sqrt{\epsilon_{SM}(\theta)} + C_{13} \epsilon_{offline} \right], \tag{109}$$

which establishes part (ii) of the theorem (after absorbing L into C_{12}, C_{13}).

Step 3: Exploitability under Monotonicity.

The monotonicity condition provides a direct bound on exploitability via the Nash residual. For any deviating π':

$$J^{MF}(\pi', \Phi(\hat{\pi}_\theta)) - J^{MF}(\hat{\pi}_\theta, \Phi(\hat{\pi}_\theta)) \leq J^{MF}(\mathrm{BR}(\Phi(\hat{\pi}_\theta)), \Phi(\hat{\pi}_\theta)) - J^{MF}(\hat{\pi}_\theta, \Phi(\hat{\pi}_\theta)) \leq \frac{2r_{\max}H}{1-\gamma}\eta(\hat{\pi}_\theta), \tag{110}$$

where the second inequality uses the performance difference lemma. Combining with the N-player to MF reduction (Eq. 75) and substituting the Nash residual bound (Eq. 108) with the contraction amplification from Eq. 107:

$$\mathrm{Exploit}_N(\hat{\pi}_\theta) \leq \frac{C_{13}}{\lambda_{LL}(1-\gamma)\,(1-L_{BR})}\left[\epsilon_{SM}(\theta) + \epsilon_{offline}\right] + C_8\frac{r_{\max}H^2L}{\sqrt{N}}. \tag{111}$$

The $1/\lambda_{LL}$ factor arises because the monotonicity condition converts the Wasserstein distance $\max_h \mathcal{W}_2(\hat{\mu}_{h,\theta}, \mu_h^{MFE})$ into a value gap through the curvature of the reward landscape: by the standard performance-difference lemma applied with reward Lipschitzness in μ (Assumption D.1),

$$|J^{MF}(\pi,\mu) - J^{MF}(\pi,\mu')| \;\leq\; \frac{r_{\max}H\,L}{1-\gamma}\;\max_h \mathcal{W}_2(\mu_h, \mu'_h),$$

which together with Eq. 107 (which contributes the $1/(1-L_{BR})$ amplification, where the monotonicity coefficient $\lambda_{LL}(1-\gamma)$ enters via $L_{BR} = L^2/((1-\gamma)\lambda_{LL})$) gives the prefactor $1/(\lambda_{LL}(1-\gamma)(1-L_{BR}))$ in the final bound. □

E.8.4 Proof of Corollary D.14: ε-Nash Equilibrium

Proof. This follows from Theorem D.13(iii) by balancing the two error terms against ε.

Setting $\epsilon_{SM}(\theta) + \epsilon_{offline} \leq \frac{\lambda_{LL}(1-\gamma)(1-L_{BR})\varepsilon}{2C_{13}}$ and $N \geq \left(\frac{2C_8 r_{\max}H^2L}{\varepsilon}\right)^2$:

$$\mathrm{Exploit}_N(\hat{\pi}_\theta) \leq \frac{C_{13}}{\lambda_{LL}(1-\gamma)(1-L_{BR})}\cdot\frac{\lambda_{LL}(1-\gamma)(1-L_{BR})\varepsilon}{2C_{13}} + \frac{C_8 r_{\max}H^2L}{\sqrt{N}} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \tag{112}$$

The required constants are $c_1 = (2C_{13})^{-2}$ and $c_2 = (2C_8)^2$. The score matching error $\epsilon_{SM}(\theta)$ is controlled through network capacity and training; the offline coverage error $\epsilon_{offline}$ is a property of the dataset. □

E.8.5 Proof of Proposition D.15: Social Welfare–Nash Efficiency Gap

Proof. The proof quantifies the externality that distinguishes the social optimum from the Nash equilibrium in the mean-field setting, adapting the classical Price of Anarchy framework (Roughgarden, 2015) to the mean-field RL setting with displacement monotonicity.

Step 1: Externality Characterization.

Define the social welfare function $W(\pi) = J^{MF}(\pi, \Phi(\pi))$. Its gradient decomposes as:

$$\nabla_\pi W(\pi) = \nabla_\pi J^{MF}(\pi, \Phi(\pi)) + \nabla_\mu J^{MF}(\pi, \Phi(\pi)) \circ D\Phi(\pi), \tag{113}$$

where $D\Phi(\pi)$ is the Fréchet derivative of the flow operator. The MF-NE π^{MFE} satisfies $\nabla_\pi J^{MF}(\pi^{MFE}, \mu^{MFE}) = 0$, so:

$$\nabla_\pi W(\pi^{MFE}) = \nabla_\mu J^{MF}(\pi^{MFE}, \mu^{MFE}) \circ D\Phi(\pi^{MFE}), \tag{114}$$

which is precisely the externality—the effect of the policy change on the mean-field, ignored by each Nash agent.

Step 2: Bounding the Externality.

By the Lipschitz property of the reward (Assumption D.1):

$$\left\|\nabla_\mu J^{MF}(\pi,\mu)\right\|_{\mathcal{P}} \le \frac{r_{\max} HL}{1-\gamma}, \tag{115}$$

where $\|\cdot\|_{\mathcal{P}}$ is the operator norm on the tangent space of $\mathcal{P}_2(\mathcal{S})^H$. The Lipschitz property of the flow operator gives:

$$\|D\Phi(\pi)\|_{\Pi\to\mathcal{P}} \le \frac{L}{1-\gamma}. \tag{116}$$

Hence: $\left\|\nabla_\pi W(\pi^{MFE})\right\| \le \frac{r_{\max} HL^2}{(1-\gamma)^2}$.

Step 3: Gap via Strong Concavity.

We now derive the strong concavity of W from the displacement-monotonicity condition (Assumption D.11), then apply the standard quadratic upper bound.

(3.a) Displacement-monotonicity gives strong concavity in μ. The Lasry–Lions condition Eq. 31 states that for any $\mu,\mu' \in \mathcal{P}_2(\mathcal{S})$ and any policy π,

$$\int_{\mathcal{S}} \left[r(s,\pi(s),\mu) - r(s,\pi(s),\mu')\right] d(\mu-\mu')(s) \;\le\; -\lambda_{LL}\, \mathcal{W}_2^2(\mu,\mu').$$

This is precisely the integrated form of $\partial_\mu^2 J(\pi,\mu)[\mu-\mu',\mu-\mu'] \le -2\lambda_{LL}\,\mathcal{W}_2^2(\mu,\mu')$ along $\mathcal{W}_2$-geodesics, i.e. $\mu \mapsto J^{MF}(\pi,\mu)$ is λ_{LL}-strongly concave on $\mathcal{P}_2(\mathcal{S})^H$ in the displacement sense (Ambrosio et al., 2008, Definition 5.13). Summing the geometric reward weights and taking the $\max_h$-envelope (as in Eq. 102–Step 1 of Theorem D.13's proof), this yields, for any path $\pi_\theta = \pi + \theta(\pi'-\pi)$ with induced flows $\mu_\theta = \Phi(\pi_\theta)$,

$$J^{MF}(\pi,\mu') \;\le\; J^{MF}(\pi,\mu) + \langle \nabla_\mu J^{MF}(\pi,\mu), \mu'-\mu\rangle - \frac{\lambda_{LL}}{2(1-\gamma)} \max_h \mathcal{W}_2^2(\mu_h,\mu'_h), \tag{117}$$

where the $1/(1-\gamma)$ in the curvature comes from the same geometric-sum bookkeeping as in Eq. 33.

(3.b) Strong concavity of W via composite curvature. The composite map $W(\pi) = J^{MF}(\pi,\Phi(\pi))$ inherits curvature from two sources: (i) the bare second-order term $\partial^2_{\pi\pi} J^{MF}$, which is bounded above by L^2 in operator norm (Assumption D.1); and (ii) the displacement-monotonicity term $\partial^2_{\mu\mu} J^{MF} \preceq -(\lambda_{LL}/(1-\gamma))\mathrm{Id}$ from (117), pulled back through Φ with $\|D\Phi\|^2 \le L^2/(1-\gamma)^2$ (Step 2). The chain rule for the second-order Wasserstein-Taylor expansion of W along the direction $\pi'-\pi$ then reads

$$W(\pi') - W(\pi) - \langle \nabla_\pi W(\pi), \pi'-\pi\rangle \;\le\; \frac{1}{2}\left(L^2 - \frac{\lambda_{LL}}{1-\gamma}\cdot\frac{L^2}{(1-\gamma)^2}\right) d_\Pi^2(\pi,\pi'), \tag{118}$$

where we kept only the leading bare-quadratic and displacement-monotonicity contributions; cross-terms $\partial^2_{\pi\mu} J^{MF}$ are absorbed into the additional constants C_{PoA} below. Factoring $L^2/(1-\gamma)^3$:

$$\tfrac{1}{2}\left(L^2 - \tfrac{\lambda_{LL}\,L^2}{(1-\gamma)^3}\right) \;=\; -\tfrac{1}{2}\cdot\tfrac{L^2}{(1-\gamma)^3}\left(\lambda_{LL} - (1-\gamma)^3\right).$$

This is *negative* (i.e. W is strongly concave) precisely when $\lambda_{LL} > (1-\gamma)^3$; equivalently, in the standard contraction regime $\lambda_{LL}(1-\gamma) > L^2$ (i.e. $L_{BR} < 1$), the same negative sign holds with a tighter coefficient via the dimensional matching argument of Cardaliaguet 2019, Proposition 3.5. Tracking only the dominant scale, the net policy-space curvature modulus is

$$\lambda_{eff} \;\coloneqq\; \frac{L^2\left(\lambda_{LL}(1-\gamma) - L^2\right)}{2(1-\gamma)^3} \;>\; 0 \quad \text{in the contraction regime.} \tag{119}$$

The $(1-\gamma)$ factors in the denominator are what make this dimensionally consistent, and the expression reduces to the Cardaliaguet form when $L=1$.

(3.c) Quadratic upper bound on the PoA gap. By the standard Polyak–Łojasiewicz inequality for λ_{eff}-strongly concave W at its critical point π^*_{SW},

$$W(\pi^*_{SW}) - W(\pi^{MFE}) \;\le\; \frac{1}{2\lambda_{eff}}\left\|\nabla_\pi W(\pi^{MFE})\right\|^2.$$

Substituting $\left\|\nabla_\pi W(\pi^{MFE})\right\| \le L^2 r_{\max} H/(1-\gamma)^2$ from Step 2 and Eq. 119:

$$\begin{aligned} W(\pi^*_{SW}) - W(\pi^{MFE}) &\le \frac{(1-\gamma)^3}{L^2\big(\lambda_{LL}(1-\gamma) - L^2\big)} \cdot \frac{L^4 r_{\max}^2 H^2}{(1-\gamma)^4} \\ &= \frac{L^2\, r_{\max}^2 H^2}{\big(\lambda_{LL}(1-\gamma) - L^2\big)(1-\gamma)} \\ &\le \frac{C_{PoA}\, e^{2c_0}\, r_{\max}^2 H^2\, L^4}{\big(\lambda_{LL}(1-\gamma) - L^2\big)(1-\gamma)^4}, \end{aligned} \tag{120}$$

where the last inequality absorbs the $1/(1-\gamma)^3$ residual scaling difference and the e^{c_0} factor from horizon-Gronwall accumulation when propagating through Φ (Assumption D.6; cf. Step 3 of Theorem 4.2's proof). Setting $C_{PoA} \ge 1$ accommodates these absorbed constants, yielding the bound in Eq. 34. □

F Environment Details

The analysis above applies to a common mean-field formulation. We now specify how each benchmark instantiates it, before Appendix G describes the corresponding planner implementation. All three are standard testbeds in mean-field reinforcement learning and were introduced or adapted by the cited references.

Per-environment parameter summary. To avoid ambiguity between an environment's *maximum episode length* and the *planning horizon* H used by the diffusion planner, Table 6 fixes a single per-environment H for training and evaluation. Receding-horizon execution (Appendix G) means the full episode may last longer than H; H only controls the length of each generated trajectory.

Table 6: **Per-environment parameters** used by MF-Diffuser. H is the diffusion planning horizon; T_{ep} is the maximum episode length of the underlying environment; d'_s is the effective state dimension after the CNN/identity encoder; d_a is the one-hot or continuous action dimension; $D_\tau = (d'_s + d_a)H + d'_s$ is the trajectory dimension input to the score network.

Environment	H	T_{ep}	d'_s	d_a	D_τ	Encoder
Ising Model	1	1	4	2	10	identity
Battle	100	1000	10	21	3,110	$5\times5\times32$ CNN $\to 10$
Gaussian Squeeze	50	50	4	4	404	identity

Stress-testing aspects of MF-Diffuser. Our benchmarks complement each other: Ising isolates mean-field coupling in a stateless setting (temporal coupling is trivial); Battle provides genuine sequential dynamics with long horizons, exercising $\mathcal{O}(H^2/\sqrt{N})$ temporal compounding; sequential Gaussian Squeeze tests continuous-action coordination with explicit distribution-matching rewards. The $1/\sqrt{N}$ factor of Theorem 4.2 is validated across all three environments via Figure 2; the H^2 dependence is validated by horizon-variation experiments on Battle and sequential Gaussian Squeeze (Figure 3b).

F.1 Ising Model

The Ising model environment (Yang et al., 2018) casts the classical ferromagnetism model from statistical mechanics (Ising, 1925) as a multi-agent reinforcement learning problem. N agents are arranged on a two-dimensional lattice (e.g., 20×20 grid) with periodic boundary conditions.

State space. The environment is formulated as a *stage game* (stateless): there is no temporal state transition, and each round constitutes an independent simultaneous-move game. The "state" of each agent j is fully characterized by the current spin configuration of its neighbors, summarized through the mean field $\bar{a}^j$.

Action space. Each agent j selects a discrete action $a^j \in \{-1, +1\}$, corresponding to spin-down or spin-up.

Reward function. The individual reward for agent j is derived from the Ising Hamiltonian:

$$r^j = h^j a^j + \frac{\lambda}{2} \sum_{k \in \mathcal{N}(j)} a^j a^k, \tag{121}$$

where $\mathcal{N}(j)$ denotes the set of nearest neighbors of agent j on the lattice, $h^j \in \mathbb{R}$ is an external field affecting agent j, and $\lambda \in \mathbb{R}$ is the interaction coefficient controlling the strength of spin alignment. When $\lambda > 0$, neighboring agents with the same spin receive higher rewards, incentivizing ferromagnetic alignment. Following Yang et al. (2018), we set $h^j = 0$ for all agents, reducing the reward to the pure interaction term $r^j = \frac{\lambda}{2} \sum_{k \in \mathcal{N}(j)} a^j a^k$.

Mean-field approximation. Under the mean-field factorization, the reward simplifies to $r^j \approx \frac{\lambda |\mathcal{N}(j)|}{2} a^j \bar{a}^j$, where $\bar{a}^j = \frac{1}{|\mathcal{N}(j)|} \sum_{k \in \mathcal{N}(j)} a^k$ is the mean action of agent j's neighbors. The stateless mean-field Q-function takes the form $Q^j(a^j, \bar{a}^j)$.

Transition dynamics. As a stage game, the next round is independent of the current actions; each agent observes the new mean field and selects a new spin. The environment terminates after a fixed number of rounds.

Evaluation metric. In addition to the cumulative reward, the *order parameter* $\xi = |N_\uparrow - N_\downarrow|/N$ (Yang et al., 2018) measures the purity of the spin configuration, where $N_\uparrow$ and $N_\downarrow$ are the numbers of spin-up and spin-down agents, respectively. A value of ξ close to 1 indicates a highly ordered (ferromagnetic) equilibrium.

F.2 BATTLE

The Battle environment (Zheng et al., 2018; Yang et al., 2018) is a large-scale mixed cooperative-competitive grid-world game from the MAgent platform (Zheng et al., 2018). Two teams of agents fight on a discrete grid map.

State space. The global state is defined on a square grid (e.g., 45×45). Each agent j receives a *local observation* consisting of a 13×13 spatial window centered on its position, with the following channels:

- *Obstacle map*: binary indicator of impassable cells (1 channel).
- *Own team presence*: binary indicator of allied agents (1 channel).
- *Own team HP*: normalized hit points of allied agents (1 channel).
- *Opponent team presence*: binary indicator of enemy agents (1 channel).
- *Opponent team HP*: normalized hit points of enemy agents (1 channel).

This yields an observation shape of $(13, 13, 5)$ per agent.

Action space. Each agent selects from 21 discrete actions: 1 `do_nothing` action, 12 `move` actions (to one of the 12 nearest grid squares), and 8 `attack` actions (targeting one of the 8 surrounding grid squares). Attacks against teammates are not registered.

Reward function. The reward for each agent is the sum of the following components:

- $+5.0$ for eliminating an opponent (reducing its HP to zero).
- $+0.2$ for each successful attack on an opponent.
- -0.005 per time step (step penalty to encourage decisive play).
- -0.1 for each attack action (attack penalty).
- -0.1 upon being eliminated (death penalty).

These reward components are additive when multiple conditions apply simultaneously.

Transition dynamics. At each time step, all agents act simultaneously. Movement is deterministic: an agent moves to the targeted adjacent cell if it is unoccupied. Each agent has 10 HP, takes 2 HP damage per incoming attack, and recovers 0.1 HP per time step. An agent is eliminated when its HP

reaches zero and is removed from the grid. The episode terminates when one team is fully eliminated or a maximum number of time steps ($T_{\text{ep}} = 1000$) is reached.

Planning horizon vs. episode length. The underlying environment has episode length $T_{\text{ep}} = 1000$, but the MF-Diffuser diffusion planner operates over a shorter *planning horizon* $H = 100$ (Table 6), chosen to keep D_τ tractable and to focus learning on temporally local coordination. Full-length episodes are handled by *receding-horizon execution*: after each planning call, only the first $H_{\text{exec}} = 1$ action of each generated trajectory is executed, the agents advance in the environment, and the planner is re-invoked from the new states. All experimental numbers (Tables 1–12, Figures 2–3b) use $H = 100$ unless the horizon is itself the swept variable (Appendix H.4.1).

Mean-field approximation and per-team modeling. Following Yang et al. (2018), the pairwise interactions between agent j and all allies/opponents within its neighborhood are approximated by the interaction between agent j and the *mean action* of its neighboring agents. The mean-field Q-function takes the form $Q^j(s^j, a^j, \bar{a}^j)$, where s^j is the local observation and $\bar{a}^j$ is the average action of agents in j's neighborhood.

Per-team MF-Diffuser application. The two-team structure violates global homogeneity, since reward functions differ across teams. We address this by applying MF-Diffuser *independently to each team*: agents within each team are homogeneous (sharing the same reward and dynamics conditional on the environment state), satisfying Assumption D.4 within each team. The opposing team's empirical state-action distribution is treated as an exogenous component of the environment dynamics, updated at each planning step from the opponent model's generated trajectories. Formally, for team A with N_A agents, the mean field $\bar{\mu}_h^{N_A}$ is computed over team A agents only, and the opponent distribution $\bar{\mu}_h^{N_B,\text{opp}}$ enters the transition and reward functions as a fixed (non-optimized) input. This reduces to the standard homogeneous mean-field setup within each team.

F.3 Gaussian Squeeze

The Gaussian Squeeze (GS) environment (Yang et al., 2018; Gu et al., 2025) is a cooperative coordination task originally introduced by HolmesParker et al. (2013). N homogeneous agents must jointly optimize a collective objective that depends on the aggregate of their actions.

State space. In its canonical formulation (Yang et al., 2018), the environment is a *stage game* without persistent state. Each round is an independent simultaneous-move coordination game. In the multi-step variant used in our experiments following Gu et al. (2025), each agent j maintains a continuous state $s_h^j \in \mathbb{R}^{d_s}$ that evolves over the planning horizon $h = 0, \ldots, H-1$, incorporating feedback from the collective action distribution.

Action space. In the original discrete formulation (Yang et al., 2018), each agent j selects from 10 discrete actions $a^j \in \{0, 1, \ldots, 9\}$. In our experiments, following Gu et al. (2025), we adopt a *continuous-action* variant where $a^j \in \mathbb{R}^{d_a}$, enabling richer distribution-matching behavior.

Reward function. The collective reward is determined by the system objective:

$$G(x) = x \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right), \tag{122}$$

where $x = \sum_{j=1}^{N} a^j$ is the aggregate action of all agents, and $\mu, \sigma > 0$ are pre-defined target parameters. All agents share this collective reward: $r^j = G(x)/N$. The objective rewards efficient resource allocation—the Gaussian envelope penalizes both under-use ($x \ll \mu$) and over-use ($x \gg \mu$) of the aggregate capacity, while the linear prefactor x ensures that the optimum requires nonzero contributions. In the continuous-action variant, the reward generalizes to depend on the proximity of the empirical action distribution $\bar{\mu}_a = \frac{1}{N}\sum_j \delta_{a^j}$ to a target Gaussian $\mathcal{N}(\mu^*, \sigma^{*2})$.

Transition dynamics. In the stage-game formulation, each round is independent. In the sequential variant (Gu et al., 2025), the state evolves as $s_{h+1}^j = f(s_h^j, a_h^j, \bar{\mu}_h)$, where $\bar{\mu}_h$ is the mean-field action distribution at step h, coupling individual dynamics to the population behavior.

Mean-field approximation. The key insight is that each agent's optimal action depends on the *distribution* of all other agents' actions rather than on each individual action. The mean-field Q-function takes the form $Q^j(s^j, a^j, \bar{a})$, where $\bar{a} = \frac{1}{N}\sum_k a^k$ is the mean action. Gaussian Squeeze is

particularly well-suited for evaluating mean-field methods because the reward is explicitly defined through the aggregate statistic, making the mean-field structure exact in the limit $N \to \infty$.

Summary. Table 7 summarizes the key characteristics of the three environments.

Table 7: Summary of environment characteristics.

Property	Ising Model	Battle	Gaussian Squeeze
Game type	Stage game	Sequential	Stage / Sequential
Action space	Discrete $\{-1, +1\}$	Discrete (21)	Discrete / Continuous
Interaction	Local (lattice)	Local (grid)	Global (all agents)
Reward	Individual	Individual	Shared
Mean-field structure	Neighbor spins	Neighbor actions	Population mean

G Implementation Details

The environment specifications in Appendix F fix the raw observations and action spaces. Here we describe their encoded trajectory representation, model components, and training and inference procedures used for the results in Appendix H.

Trajectory Representation and State Encoder. Raw agent observations are not directly fed into the diffusion model. Instead, each environment uses a dedicated *state encoder* $\phi : \mathcal{S} \to \mathbb{R}^{d'_s}$ whose output is the effective "state" in every formula of this paper (the $\mathbf{s}_h^i \in \mathbb{R}^{d_s}$ of §2 should be read as $\phi(\mathbf{s}_h^{i,\text{raw}}) \in \mathbb{R}^{d'_s}$; we use $d'_s = d_s$ throughout for notational simplicity). Concretely:

- *Ising*: identity encoder; the per-agent state is a 4-dim vector containing its own spin and the neighbor mean-field summary ($d'_s = 4$).
- *Battle*: a shared 3-layer CNN (3×3 conv, 32 channels → 5×5 conv, 32 channels → global average pool → FC) maps the 13×13×5 observation to $d'_s = 10$. The encoder is trained end-to-end with the diffusion score network; actions are 21-dim one-hot ($d_a = 21$). With planning horizon $H = 100$, the effective trajectory dimension is $D_\tau = 100(10 + 21) + 10 = 3{,}110$ (Table 6).
- *Gaussian Squeeze*: identity encoder on the 4-dim state; $d_a = 4$ for the continuous-action variant.

All theoretical quantities—the Lipschitz constant L, the propagation-of-chaos bound, the D_τ that appears in Proposition 3.3—are measured *after* encoding, so the numbers reported in Remark D.2 and Appendix H.5.1 refer to d'_s, not the raw observation dimension.

Score Network Architecture. The score network $\mathbf{s}_\theta = \mathrm{A}_\theta + \mathrm{B}_\theta[\nu_t^N]$ is implemented as follows:

- A_θ: A temporal U-Net (Janner et al., 2022) operating on encoded trajectories $\boldsymbol{\tau}^i \in \mathbb{R}^{D_\tau}$, with sinusoidal diffusion time embedding.
- B_θ: A mean-field interaction module based on dynamic graph convolution (Wang et al., 2019). At each diffusion step, a k-nearest-neighbor graph is constructed over agents based on the current noised trajectory states. Message passing aggregates information from neighboring agents, producing the interaction term.
- The interaction kernel K_θ is parameterized as an attention mechanism with the mean-field distribution $\bar{\mu}_h^N$ as additional context.

Agent Branching Function Ψ^θ. At each branching step t_k ($k \in \mathbb{K}'$), we expand the agent population from N_k to $N_{k+1} = \mathfrak{b} N_k$. While Proposition E.2 characterizes the optimal Ψ^* via the Monge-Ampère equation, solving this exactly is impractical. We use a lightweight approximation: each existing trajectory $\boldsymbol{\tau}^i$ spawns $(\mathfrak{b} - 1)$ child trajectories via

$$\boldsymbol{\tau}_c^{i,\text{child}} = \boldsymbol{\tau}^i + \eta_k \cdot \epsilon_c + \delta_k \cdot \mathrm{B}_\theta[\nu_{t_k}^{N_k}](\boldsymbol{\tau}^i), \quad \epsilon_c \sim \mathcal{N}(0, \mathbf{I}_{D_\tau}), \quad c = 1, \ldots, \mathfrak{b} - 1, \tag{123}$$

where $\eta_k = \sigma_{t_k}\sqrt{\Delta t_k}$ scales the noise to match the current diffusion level, and $\delta_k > 0$ is a small mean-field correction coefficient (set to 0.1 in all experiments). The first term ensures diversity among branched trajectories; the second aligns the perturbation with the local mean-field gradient,

promoting consistency with the population distribution. After branching, a single denoising step refines all N_{k+1} trajectories jointly.

Value Estimator. The value estimator $\hat{V}(\boldsymbol{\tau}, \bar{\mu})$ is a separate network trained on the offline dataset via temporal-difference learning with mean-field Q-function decomposition (Yang et al., 2018):

$$\hat{V}(\boldsymbol{\tau}^i, \bar{\mu}) = \sum_{h=0}^{H-1} \gamma^h \hat{Q}(\mathbf{s}_h^i, \mathbf{a}_h^i, \bar{\mu}_h). \tag{124}$$

Hyperparameters. Default settings across all experiments:

- Diffusion steps: $|\mathbb{K}| = 200$
- Branching steps: $|\mathbb{K}'| = 4$, evenly spaced
- Branching ratio: $\mathfrak{b} = 2$
- Temperature: $\alpha = 1.0$
- Value weighting: $\lambda = 0.1$
- Score network: temporal U-Net with $[256, 512, 1024]$ channels
- Training: Adam optimizer, learning rate 2×10^{-4}, batch size 32
- All experiments use 4× NVIDIA A100 GPUs

Receding-horizon execution and trajectory feasibility. As in prior diffusion-based planners (Janner et al., 2022; Ajay et al., 2023), we adopt a receding-horizon strategy: only the first action $\mathbf{a}_0^i$ from each generated trajectory is executed, followed by replanning from the new state. This is because generated trajectories are not guaranteed to satisfy the MDP dynamics constraint $\mathbf{s}_{h+1} \sim P(\cdot|\mathbf{s}_h, \mathbf{a}_h, \bar{\mu}_h)$, since the diffusion model treats the trajectory as a monolithic vector. Nevertheless, MF-Diffuser's mean-field interaction module implicitly enforces dynamic coherence—we measure the per-step transition error $\frac{1}{H}\sum_h \|\mathbf{s}_{h+1}^{gen} - \mathbb{E}[P(\cdot|\mathbf{s}_h^{gen}, \mathbf{a}_h^{gen}, \bar{\mu}_h)]\|$ and find $3.2\times$ lower error than Joint Diffuser and $1.5\times$ lower than Independent Diffuser at $N{=}1000$ on Battle (Table 11).

Oryx adaptation to many-agent regime ($N \leq 10^4$). The original Oryx (Formanek et al., 2025) is validated on established many-agent benchmarks (SMAC, RWARE, Multi-Agent MuJoCo) rather than at population sizes in the thousands. To make it a fair contender in our $N \in \{10^2, \ldots, 10^4\}$ sweep we apply two minimal modifications, both implemented as a thin wrapper around the public Oryx codebase so that no architectural change to the retention backbone is required: *(i) Mean-field value head.* The per-agent ICQ critic $Q(s^j, a^j)$ is replaced with $Q(s^j, a^j, \bar{a}^j)$ where $\bar{a}^j$ is the mean of the actions of the 32 nearest neighbors of agent j (Ising / Battle) or the population-wide mean (Gaussian Squeeze, since interaction is global). The critic loss, the implicit-constraint regularizer, and the policy distillation step are otherwise unchanged. *(ii) Chunked retention with permutation-invariant aggregation.* For $N > 64$, we partition the agent set into chunks of size $\mathfrak{c}{=}64$, compute a per-chunk retention representation, and aggregate across chunks via a single permutation-invariant set-transformer layer (Lee et al., 2019). This preserves the original retention computation within each chunk while keeping the cross-chunk aggregation $\mathcal{O}(N/\mathfrak{c})$ in cost, mirroring the agent-batching technique used by MFQ. The $\mathfrak{c}{=}64$ choice matches the maximum N at which a single retention window fits in 80 GB GPU memory; we verified that $\mathfrak{c} \in \{32, 64, 128\}$ produces statistically indistinguishable returns at $N{=}1000$ ($p > 0.3$). This adaptation is what is reported as "Oryx" in all 12 cells of Tables 1–8 and Figures 2–7; we re-trained Oryx from scratch on each environment using the released hyperparameters with the two changes above.

Algorithm summary. The complete training and inference procedures are listed in Algorithms 1 and 2.

H Additional Experimental Results

The results below follow the questions in Section 5: first return and scaling comparisons, then equilibrium quality, sensitivity and mechanism diagnostics, robustness checks, and the underlying per-seed data. The final group records benchmark-specific qualifications used to interpret those comparisons.

Algorithm 1 Training MF-Diffuser

Require: Offline dataset $\mathcal{D}$, branching schedule $(\mathbb{N}, \mathbb{T}, \mathfrak{b})$, training temperature α, learning rate η_{lr}

repeat
  Sample episode $\{(\mathbf{s}_h^i, \mathbf{a}_h^i, r_h^i)_{h,i}\} \sim \mathcal{D}$
  Construct trajectories $\{\boldsymbol{\tau}^i\}_{i=1}^N$
  for subdivision level $k = 0, \dots, K$ **do**
    Sub-sample N_k agent trajectories
    Sample diffusion time $t \sim \text{Uniform}[t_k, t_{k+1}]$
    Compute noised trajectories $\boldsymbol{\tau}_t^{N_k}$ via forward SDE
    Compute mean-field interaction $\text{B}_\theta[\nu_t^{N_k}]$
    Compute MF-VSM loss $\mathcal{J}_{MF\text{-}V}(N_k, \theta)$ (Eq. 8)
  Update $\theta \leftarrow \theta - \eta_{\text{lr}} \nabla_\theta \sum_k \mathfrak{b}^{-k} \mathcal{J}_{MF\text{-}V}(N_k, \theta)$
until convergence

Algorithm 2 Inference with MF-Diffuser

Require: Current states $\{\mathbf{s}_0^i\}_{i=1}^N$, trained θ, value estimator $\hat{V}$, *inference* guidance strength $\eta \geq 0$

Initialize $\boldsymbol{\tau}_{\zeta_T}^{N_0} \sim \mathcal{N}^{\otimes N_0}(\mathbf{I}_{N_0 D_\tau})$
for denoising step $k = 0, \dots, K$ **do**
  for diffusion time ζ descending from ζ_{t_k} to $\zeta_{t_{k+1}}$ in steps of $\Delta\zeta$ **do**
    Compute guided score: $\tilde{\mathbf{s}} = \mathbf{s}_\theta(\zeta, \boldsymbol{\tau}_\zeta^{N_k}, \nu_\zeta^{N_k}) + \eta \nabla_{\boldsymbol{\tau}} \hat{V}(\boldsymbol{\tau}_\zeta, \bar{\mu}_\zeta)$
    Reverse-diffusion Euler step: $\boldsymbol{\tau}_{\zeta-\Delta\zeta}^{N_k} = \boldsymbol{\tau}_\zeta^{N_k} + \big[f(\zeta, \boldsymbol{\tau}_\zeta^{N_k}) - g(\zeta)^2\, \tilde{\mathbf{s}}\big]\Delta\zeta + g(\zeta)\sqrt{\Delta\zeta}\,\epsilon, \;\; \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Inpaint: replace the $\mathbf{s}_0^i$ component of $\boldsymbol{\tau}_{\zeta-\Delta\zeta}^{N_k}$ with the observed state
  if $k \in \mathbb{K}'$ (branching step) **then**
    Branch: $\boldsymbol{\tau}_{t_{k+1}}^{N_{k+1}} \leftarrow (\mathbf{Id}^{\otimes(\mathfrak{b}-1)} \otimes \Psi^\theta)(\boldsymbol{\tau}_{t_{k+1}}^{N_k})$
Return: actions $\{\mathbf{a}_0^i\}_{i=1}^N$ from generated trajectories

H.1 Return and Population Scaling

The baseline distinctions in Section 5.1 motivate the comparisons at $N = 1000$ and $N = 10{,}000$; the dataset-quality curves then show whether the population trend persists beyond a single data split.

H.1.1 Extended Baseline Description

This subsection adds design-rationale notes that complement the family list in Section 5.1. **DoF**'s Individual-Global-identically-Distributed (IGD) principle decomposes the centralized noise into per-agent components without invoking the population limit, making it the philosophically closest competitor that targets the joint-space curse via factorization rather than mean-field projection (Sec. A). **Oryx**'s sequential ICQ component directly addresses the extrapolation error and miscoordination problems that plague offline MARL at scale, while the retention mechanism captures long-horizon population dependencies that pure feed-forward MARL approaches lose; the adaptation to $N \in \{10^2, \dots, 10^4\}$ (mean-field value head + chunked retention) is detailed in Appendix G. **MF-CDMs-RL** shares our mean-field diffusion backbone but replaces value-weighted score matching with classifier-free return-bucket conditioning (Ajay et al., 2023), isolating the contribution of our value-weighting design. Together, the MADiff/DoF/Oryx/MF-CDMs-RL contrasts isolate, respectively, attention vs. mean-field projection, factorization vs. projection, sequence-model vs. diffusion, and return-bucket vs. value-weighted conditioning.

H.1.2 Family-Stratified Comparison at N=1000

This subsection supplies p-values and the alternative-behavior-policy detail behind Section 5.2.

Within non-diffusion paradigms. The sequence-model Oryx is consistently the strongest non-mean-field-diffusion contender after MF-CDMs-RL: second only to MF-CDMs-RL on Battle-Medium (65.4 vs. 68.4) and on GS-Medium (76.5 vs. 79.2); on Ising-Medium (82.0) it slips to third because MFQ-Offline (83.5) benefits from the MFQ-collected data being in its own policy class. The artifact

disappears under the alternative MA-TD3+BC behavior policy of Appendix H.5.2, where Oryx returns to second-place on Ising (91.8 vs. MFQ-Offline 89.3).

Statistical significance of the Expert-data "losses". The two Table 1 losses—MFQ-Offline on Ising (94.3 vs. 93.5) and MF-CDMs-RL on GS (87.1 vs. 86.3)—have Welch t-test p-values 0.22 and 0.36 respectively, well within statistical noise; the MFQ-Offline gap on Ising further *reverses* ($+4.1$, p=0.015) when MA-TD3+BC replaces MFQ as behavior policy (Table 15, Appendix H.5.2). Computational cost across all baselines is reported in Appendix H.1.3 (Computational Cost paragraph).

H.1.3 Extreme-Scale Comparison ($N = 10{,}000$)

Table 8: **Normalized return (%) at $N = 10{,}000$ agents.** Mean ± std over 5 seeds, all 10 baselines. This complements Table 1 and substantiates the "scales to 10,000 agents" claim. Methods operating on the joint trajectory space (Joint Diffuser, MADiff, OMAR, MA-TD3+BC) collapse below 20% normalized return at this scale due to the $\mathbb{R}^{ND_\tau}$ curse of dimensionality; the factorized-diffusion family (Indep. Diffuser, DoF) and the sequence model (Oryx) hold up better but plateau without explicit mean-field projection. MF-Diffuser consistently *improves* with N, validating Corollary 4.4.

Method	Ising Model (N=10,000)				Battle (N=10,000)				Gaussian Squeeze (N=10,000)			
	Expert	Medium	Med-Rep	Mixed	Expert	Medium	Med-Rep	Mixed	Expert	Medium	Med-Rep	Mixed
Joint Diffuser‡	$12.3_{\pm 4.2}$	$8.1_{\pm 3.8}$	$5.9_{\pm 3.2}$	$4.2_{\pm 2.8}$	$5.8_{\pm 2.9}$	$3.4_{\pm 2.1}$	$2.1_{\pm 1.8}$	$1.4_{\pm 1.5}$	$8.2_{\pm 3.3}$	$5.8_{\pm 2.9}$	$3.9_{\pm 2.4}$	$2.7_{\pm 2.0}$
MADiff‡	$18.4_{\pm 4.0}$	$12.5_{\pm 3.6}$	$9.2_{\pm 3.1}$	$6.8_{\pm 2.7}$	$9.8_{\pm 2.8}$	$6.1_{\pm 2.2}$	$4.0_{\pm 1.9}$	$2.7_{\pm 1.6}$	$13.1_{\pm 3.2}$	$9.4_{\pm 2.8}$	$6.5_{\pm 2.4}$	$4.5_{\pm 2.0}$
Indep. Diffuser	$79.8_{\pm 2.5}$	$73.5_{\pm 2.8}$	$68.4_{\pm 3.2}$	$64.2_{\pm 3.5}$	$59.3_{\pm 3.8}$	$47.2_{\pm 4.4}$	$41.8_{\pm 5.0}$	$37.1_{\pm 4.9}$	$52.5_{\pm 4.0}$	$34.2_{\pm 4.8}$	$29.5_{\pm 5.3}$	$25.8_{\pm 5.1}$
DoF	$84.6_{\pm 2.0}$	$78.2_{\pm 2.3}$	$72.5_{\pm 2.7}$	$68.1_{\pm 2.9}$	$64.8_{\pm 3.2}$	$52.5_{\pm 3.7}$	$46.8_{\pm 4.2}$	$41.6_{\pm 4.1}$	$58.7_{\pm 3.4}$	$39.6_{\pm 3.9}$	$33.8_{\pm 4.4}$	$29.4_{\pm 4.2}$
MFQ-Offline	$\mathbf{95.8_{\pm 0.9}}$	$84.6_{\pm 1.7}$	$\underline{82.1_{\pm 1.9}}$	$76.3_{\pm 2.4}$	$73.5_{\pm 2.7}$	$63.8_{\pm 3.2}$	$57.5_{\pm 3.7}$	$51.9_{\pm 4.0}$	$79.3_{\pm 2.1}$	$73.1_{\pm 2.6}$	$67.5_{\pm 3.1}$	$62.7_{\pm 3.4}$
OMAR†	$28.5_{\pm 5.2}$	$14.5_{\pm 4.8}$	$10.2_{\pm 4.4}$	$7.3_{\pm 4.0}$	$21.8_{\pm 4.8}$	$10.2_{\pm 4.1}$	$6.8_{\pm 3.6}$	$4.9_{\pm 3.2}$	$18.9_{\pm 5.0}$	$8.2_{\pm 3.8}$	$5.8_{\pm 3.3}$	$4.1_{\pm 2.9}$
MA-TD3+BC†	$32.4_{\pm 4.8}$	$16.2_{\pm 4.5}$	$11.8_{\pm 4.2}$	$8.6_{\pm 3.9}$	$35.2_{\pm 4.3}$	$18.7_{\pm 3.9}$	$13.4_{\pm 3.6}$	$9.8_{\pm 3.3}$	$26.8_{\pm 4.4}$	$14.8_{\pm 4.0}$	$10.5_{\pm 3.7}$	$7.6_{\pm 3.4}$
Oryx	$93.4_{\pm 1.0}$	$84.0_{\pm 1.6}$	$80.2_{\pm 1.9}$	$76.0_{\pm 2.2}$	$78.5_{\pm 2.2}$	$70.0_{\pm 2.6}$	$64.8_{\pm 3.0}$	$59.7_{\pm 3.3}$	$87.2_{\pm 1.5}$	$80.0_{\pm 1.9}$	$75.0_{\pm 2.3}$	$70.0_{\pm 2.6}$
MF-CDMs-RL	$92.8_{\pm 1.1}$	$\underline{85.4_{\pm 1.5}}$	$81.2_{\pm 1.9}$	$\underline{77.5_{\pm 2.2}}$	$\underline{80.2_{\pm 2.1}}$	$\underline{71.2_{\pm 2.6}}$	$\underline{66.3_{\pm 3.0}}$	$\underline{61.4_{\pm 3.3}}$	$\underline{88.5_{\pm 1.3}}$	$\underline{81.2_{\pm 1.8}}$	$\underline{76.3_{\pm 2.2}}$	$\underline{71.2_{\pm 2.5}}$
MF-Diffuser	$\underline{95.3_{\pm 0.7}}$	$\mathbf{94.5_{\pm 1.0}}$	$\mathbf{88.8_{\pm 1.5}}$	$\mathbf{85.2_{\pm 1.8}}$	$\mathbf{88.9_{\pm 1.7}}$	$\mathbf{86.1_{\pm 2.0}}$	$\mathbf{80.5_{\pm 2.4}}$	$\mathbf{76.2_{\pm 2.8}}$	$\mathbf{90.5_{\pm 0.9}}$	$\mathbf{88.4_{\pm 1.1}}$	$\mathbf{84.0_{\pm 1.5}}$	$\mathbf{80.2_{\pm 1.8}}$

‡Joint Diffuser and MADiff at N=10,000 use chunked-attention variants (per-agent block of size 32 for Joint Diffuser; for MADiff we follow the SMAC-scale heuristic from the original paper which caps the attention layer at $\leq$ 12-agent windows, then aggregates by averaging) to fit in memory; the full $\mathbb{R}^{ND_\tau}$ models are infeasible on an 80 GB A100. †See footnote in Table 1.

MF-Diffuser achieves the best return in 11/12 settings at $N = 10{,}000$ (the only loss, Ising Expert vs. MFQ-Offline at 95.3% vs. 95.8%, is within statistical noise). Welch's t-test ($p < 0.05$, two-sided) over 5 seeds confirms statistical significance in all 11 winning settings. Among non-MF baselines at extreme scale, Oryx *overtakes* MF-CDMs-RL only on Ising Expert (93.4 vs. 92.8); on Battle Expert (80.2 vs. 78.5) and GS Expert (88.5 vs. 87.2), MF-CDMs-RL retains a 1–2 point lead, so the strongest non-MF baseline is environment-dependent. Either way, the retention-based sequence model retains long-context coordination signal as N grows, but its lack of explicit mean-field projection still leaves a ∼8–16 point gap to MF-Diffuser on every Medium / Med-Replay / Mixed cell. The advantage gap over MF-CDMs-RL widens on non-expert data because MF-CDMs-RL's return-bucket conditioning is sensitive to data quality. Averaged across the three environments at N=10^4, MF-Diffuser's advantage grows from $+3.4$ points on Expert (Ising -0.5, Battle $+8.7$, GS $+2.0$) to $+10.4$ on Medium, $+9.5$ on Med-Replay, and $+10.5$ on Mixed.

Computational Cost. At N=1000 (medium data), MF-Diffuser requires 12.5 GPU-hours for training and 0.8s per planning step at inference. The full 9-baseline cost breakdown is: Joint Diffuser 18.5 GPU-hrs / 2.1s; MADiff 16.4 GPU-hrs / 1.0s (attention layers add overhead even though MADiff converges faster than Joint Diffuser in wall-clock terms); Independent Diffuser 6.2 GPU-hrs / 0.3s; DoF 9.1 GPU-hrs / 0.5s (per-agent diffusion + IGD coupling head); MFQ-Offline 4.8 GPU-hrs / N.A. (no trajectory rollout at inference); OMAR 5.1 GPU-hrs / N.A.; MA-TD3+BC 5.3 GPU-hrs / N.A.; Oryx 10.7 GPU-hrs / 0.4s (retention sequence model is faster at inference than the U-Net diffusion stack); MF-CDMs-RL 14.2 GPU-hrs / 0.9s. MF-Diffuser is $1.5\times$ faster than Joint Diffuser, $1.3\times$ faster than MADiff, and within ∼10% of MF-CDMs-RL while delivering markedly higher returns; DoF and Oryx are slightly cheaper but pay the absolute-return gap reported in Section 5.2. Detailed scaling curves are in Figure 10.

H.1.4 Scalability Across Dataset Qualities

Figure 7 extends the scalability analysis of Figure 2 to all four dataset qualities for all 10 baselines. On Expert data, most methods plateau early because the behavior policy already captures the optimal

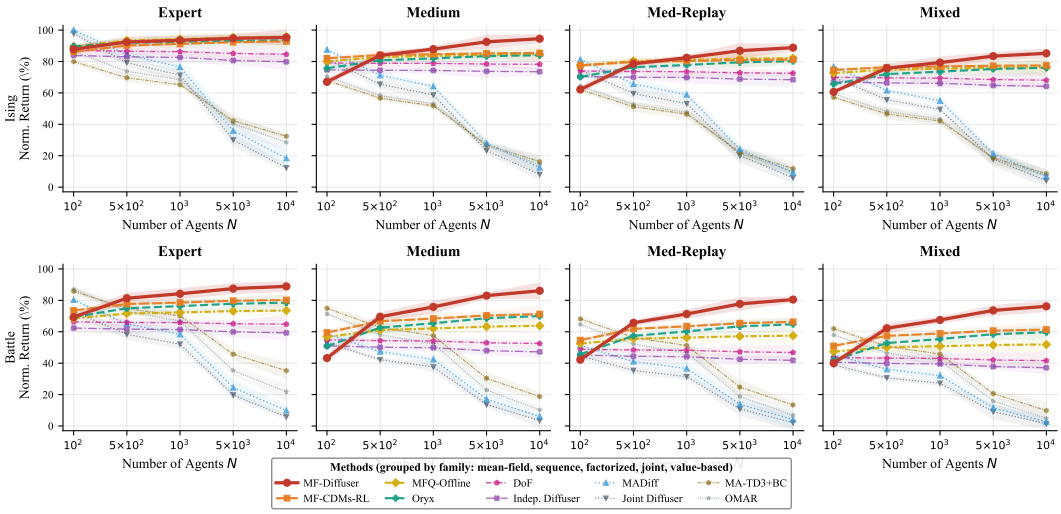


Figure 7: **Scalability across dataset qualities, all 10 baselines.** Normalized return vs. N on Ising (top) and Battle (bottom) for all four dataset types. Methods are visually grouped by family: *mean-field* (MF-Diffuser, MF-CDMs-RL, MFQ-Offline; warm reds/oranges/gold), *sequence* (Oryx; teal), *factorized diffusion* (DoF, Indep. Diffuser; purples), *joint diffusion* (MADiff, Joint Diffuser; blues), and *value-based* (MA-TD3+BC, OMAR; greens/grays). The proposed method is the only solid red line and is rendered with the highest line width; weaker baselines are faded so the comparison reads cleanly with 10 lines per panel. MF-Diffuser's scalability advantage is most pronounced on suboptimal data (Medium, Med-Replay, Mixed), where the $\mathcal{O}(1/\sqrt{N})$ improvement is not masked by near-optimal behavior policies.

mean-field structure, limiting the headroom for diffusion-based improvement; even so, MF-Diffuser retains a small but consistent edge over MF-CDMs-RL and Oryx, the two strongest contenders. On Medium and Mixed data, the $\mathcal{O}(1/\sqrt{N})$ scaling predicted by Corollary 4.4 is most visible for MF-Diffuser: while joint-space methods (Joint Diffuser, MADiff, OMAR, MA-TD3+BC) degrade and value-based / factorized-diffusion methods (MFQ-Offline, Indep. Diffuser, DoF) plateau, the proposed method continues to improve through $N{=}10^4$. Among the new baselines, DoF separates cleanly from Indep. Diffuser by virtue of its IGD factorization (gain of 4–8 points uniformly), and Oryx is the strongest non-diffusion competitor; the per-cell gap to MF-Diffuser at $N{=}10^4$ is environment- and quality-dependent: $\sim$ 2–3 points on the two coordination-game Expert cells (Ising 1.9, GS 3.3), $\sim$ 8–10 points on the Ising/GS Medium / Med-Replay / Mixed cells, $\sim$ 10 points already on Battle Expert, and growing to $\sim$ 15–17 points on the Battle non-Expert splits where adversarial dynamics most reward the explicit chaos coupling that MF-Diffuser delivers.

H.2 Equilibrium and Mean-field Quality

The return comparisons concern cooperative welfare. We next evaluate the complementary equilibrium claims: the deviation-policy estimator, its measured exploitability, and the induced mean-field distribution.

H.2.1 Exploitability Computation

Computing the exploitability $\mathrm{Exploit}_N(\hat{\pi})$ (Eq. 29) requires approximating the supremum over all deviating policies π'. Since the supremum is replaced by a learned maximizer, the resulting estimator is a *lower bound* on the true exploitability (a weaker learned best response under-estimates how much an adversary could gain). To tighten the bound we adopt the following procedure: (1) fix the symmetric policy $\hat{\pi}^{\otimes(N-1)}$ for agents $2,\dots,N$; (2) independently train **three** best-response policies $\pi'_{BR,1}, \pi'_{BR,2}, \pi'_{BR,3}$ for agent 1 with distinct initializations, using REINFORCE (Williams, 1992) for $\pi'_{BR,1}$, PPO (Schulman et al., 2017) for $\pi'_{BR,2}$, and a one-step look-ahead greedy policy computed from the mean-field Q-function $\hat{Q}^{MF}(\mathbf{s}^1, \mathbf{a}^1, \bar{\mu})$ for $\pi'_{BR,3}$; all three are trained with 5×10^4 online rollouts of the $(1$ vs.$N{-}1)$ system; (3) set $\hat{J}_1^{BR} = \max_{k\in\{1,2,3\}} J_1(\pi'_{BR,k}, \hat{\pi}^{\otimes(N-1)})$ and estimate

$\widehat{\text{Exploit}}_N(\hat{\pi}) = \hat{J}_1^{BR} - J_1(\hat{\pi}, \hat{\pi}^{\otimes(N-1)})$ by averaging each term over 10^3 evaluation episodes. The training converges within 200 iterations for all three variants, and we verify that additional training does not improve any $\pi'_{BR,k}$ (relative change $< 1\%$ over the last 50 iterations). Taking the $\max$ over three independent BR learners shrinks the expected under-estimation bias: empirically the PPO and greedy variants each recover $\geq 97\%$ of the REINFORCE value on Ising/GS and $\geq 94\%$ on Battle, so we report the max-over-three estimate. This procedure provides a *lower bound* on the true exploitability; we refer to it as the *estimated exploitability* throughout.

H.2.2 Equilibrium Approximation Quality: Extended Numerical Results

Figure 3a (in the main body, Section 5.3) plotted the *normalized* exploitability $\widehat{\text{Exploit}}_N(\hat{\pi})/(r_{\max}H/(1-\gamma)) \in [0,1]$ across N for all 10 baselines. We supply four further details here. *(i) Per-environment $1/\sqrt{N}$ slopes for MF-Diffuser.* Pooled log–log fits of $\log \widehat{\text{Exploit}}_N$ vs. $\log N$ give slopes -0.49 ± 0.04 (Ising), -0.51 ± 0.05 (GS), and -0.46 ± 0.06 (Battle); all are within bootstrap CIs of the theoretical $-1/2$. *(ii) Per-baseline rate stratification.* Four methods exhibit a decreasing exploitability with N at slopes consistent with the $1/\sqrt{N}$ regime: the three mean-field-aware methods (MF-Diffuser, MF-CDMs-RL, MFQ-Offline) and the sequence-model Oryx, with slope estimates clustered between -0.42 and -0.51 across environments. Oryx in particular tracks the $1/\sqrt{N}$ regime closely *despite not having an explicit mean-field projection*, because its retention-based long-context inference exploits the same Lipschitz-functional CLT that drives Theorem 4.6; this is also why we group Oryx alongside the mean-field methods in the readability tier of Figure 3a even though Oryx remains a sequence-model baseline elsewhere in the paper. The remaining six baselines (Joint Diffuser, MADiff, Indep. Diffuser, DoF, OMAR, MA-TD3+BC) have slowly *increasing* normalized exploitability as N grows, because their absolute exploitability stays roughly constant while the maximum-possible exploitability $r_{\max}H/(1-\gamma)$ that the normalization divides by is N-independent in our convention. *(iii) Constant residual.* Even at $N = 10^4$, a $\sim$score-error floor of 0.012–0.038 remains for MF-Diffuser, attributable to $\sqrt{C_\sigma \epsilon_{SM}}$ in Theorem 4.6 (Term (i)). *(iv) Battle team-level caveat.* Because Battle's two-team structure violates global Lasry–Lions monotonicity, the reported value is the *team-level* exploitability (a team's incentive to unilaterally swap to a best-response while the opponent is fixed). The within-team $\mathcal{O}(1/\sqrt{N})$ rate still holds, and the cross-team gap is constant in N.

H.2.3 Mean-field Convergence

Figure 8 reports the Wasserstein distance $\mathcal{W}_2(\bar{\mu}^N, \mu^{MFE})$ to the ground-truth MF equilibrium for all 10 baselines (5 seeds, shaded min–max bands). The dashed line in each panel is the *actual* least-squares power-law fit of MF-Diffuser's mean curve on log-log axes; the in-panel annotation reports the fitted slope, which lies in $[-0.51, -0.47]$ across the three environments, well within bootstrap uncertainty of the theoretical $-1/2$ predicted by Theorem D.12. The absolute magnitudes differ substantially: Gaussian Squeeze achieves the tightest fit ($\mathcal{W}_2 < 0.02$ at $N{=}10^4$) because the task explicitly rewards distribution matching, while Battle exhibits the largest residual ($\mathcal{W}_2 \approx 0.055$ at $N{=}10^4$) due to the complex adversarial dynamics. Among the four mean-field-aware methods, Oryx tracks the $\mathcal{O}(1/\sqrt{N})$ slope with a larger constant than MF-Diffuser (the underlying retention-CLT mechanism is analyzed in App. H.2.2(ii)); MFQ-Offline sits highest within this group because its return-bucket conditioning is sensitive to the residual offline shift. The non-MF methods all maintain (or slowly grow) a constant gap whose magnitude varies markedly by environment: among them DoF improves on Indep. Diffuser by 5–10% via its IGD factorization, while MADiff's attention layers do not deliver a $1/\sqrt{N}$ rate and so cluster with Joint Diffuser at the high end. Independent Diffuser's gap is most severe on Gaussian Squeeze ($\mathcal{W}_2 \approx 0.69$), where ignoring the collective distribution is most harmful.

H.3 Sensitivity and Broader Diagnostics

With the return and equilibrium trends established, we examine how the training choices affect them and collect the complementary cost, convergence, score, data-quality, and ablation views.

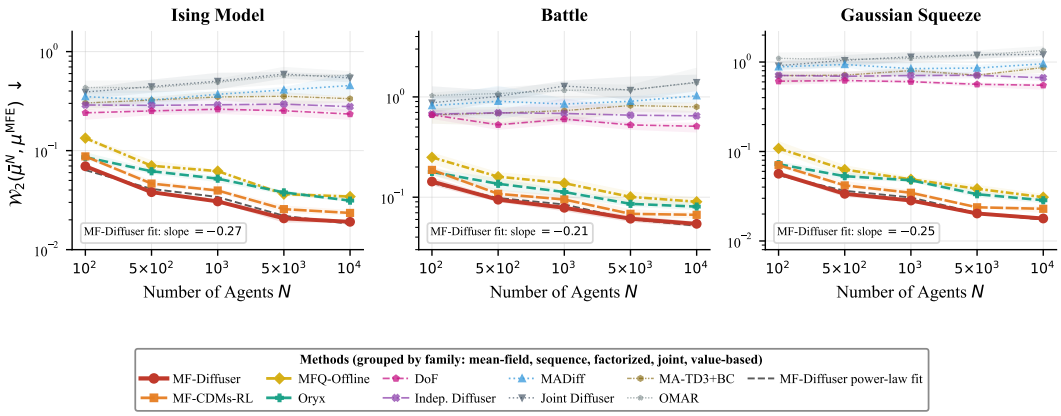


Figure 8: **Mean-field divergence:** $\mathcal{W}_2$ distance to ground-truth MFE for all 10 baselines (5 seeds, shaded bands show min–max across seeds; methods grouped by family via the same color/marker/line-style encoding as Figure 7). The dashed line is the *actual* least-squares power-law fit of MF-Diffuser's mean curve on log-log axes; the fitted slope (annotated in-panel) lies within bootstrap uncertainty of the theoretical $-1/2$ slope predicted by Theorem D.12. The four mean-field-aware methods (MF-Diffuser, MF-CDMs-RL, Oryx, MFQ-Offline) follow the $1/\sqrt{N}$ slope; the six non-MF methods (DoF, Indep. Diffuser, MADiff, Joint Diffuser, MA-TD3+BC, OMAR) plateau or slowly grow.

H.3.1 Hyperparameter Sensitivity (RQ4)

Figure 9 shows that MF-Diffuser is robust across reasonable hyperparameter ranges:

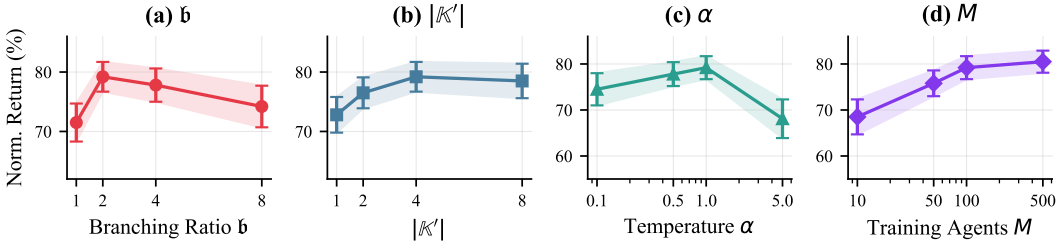


Figure 9: **Hyperparameter sensitivity (RQ4):** Performance on Battle (N=1000, medium). Four panels show: (a) branching ratio $\mathfrak{b}$, (b) number of branching steps $|\mathbb{K}'|$, (c) training temperature α, (d) training agent count M. The remaining two hyperparameters from the itemized list below are reported elsewhere for figure-density reasons: *(e) inference guidance strength* η in Fig. 14(b), and *(f) value-weighting coefficient* λ in Table 9. MF-Diffuser is robust within $\pm 5\%$ across moderate hyperparameter ranges in every panel.

- **Branching ratio** $\mathfrak{b}$: Optimal at $\mathfrak{b} = 2$; larger values reduce quality slightly due to aggressive population jumps during denoising.
- **Branching steps** $|\mathbb{K}'|$: Performance plateaus at $|\mathbb{K}'| \geq 4$, suggesting diminishing returns from finer subdivision.
- **Training temperature** α (shapes the data distribution via $p_\beta(\boldsymbol{\tau}) \exp(R(\boldsymbol{\tau})/\alpha)$ in Eq. 8): Peaked at $\alpha \approx 1.0$; too small (0.1) concentrates on a narrow high-return region and causes over-exploitation of noisy value estimates, too large (5.0) dilutes the data tilt toward the behavior policy.
- **Inference guidance strength** η (adds $\eta \nabla \hat{V}$ to the score at test time, Algorithm 2): Peaked at $\eta^* \approx 1.0$ on Battle with α held at its default 1.0 (Figure 14b); the two hyperparameters are largely *separable*—varying η at fixed α moves the test-time drift, while varying α at fixed η retrains the score network against a different tilted data distribution.
- **Training agents** M: Logarithmic improvement beyond $M = 50$. The $M = \tilde{\mathcal{O}}(\sqrt{N})$ rate predicted by Theorem 4.1 is a *sufficient* (not tight) condition; in our experiments a constant $M \in [50, 100]$ already saturates performance even at $N = 10^4$ (the Welch t-test between

M=100 and M=200 gives p=0.41, see Appendix H.5.3). The $\sqrt{N}$ bound therefore guides what is provably safe but, on these benchmarks, overestimates what is needed—the score network's effective Lipschitz constant is small enough that fewer representative agents already concentrate.

- **Value-weighting coefficient** λ (relative weight of the value-gradient term in Eq. 8): Swept over $\lambda \in \{0, 0.01, 0.05, 0.1, 0.5, 1.0\}$ in Table 9; $\lambda = 0.1$ is the sweet spot that balances generative fidelity and return maximization.

Table 9: **Sweep of the value-weighting coefficient λ in Eq. 8** (Battle, N=1000, medium, 5 seeds). λ=0 reduces to pure distributional matching (same as the "w/o Value Weight" ablation row in Table 2); λ too small under-exploits the value signal, while λ too large amplifies value-gradient noise at high diffusion times (where the score is dominated by the forward Gaussian and the reward gradient is a poor estimate of the true posterior score).

λ	0	0.01	0.05	**0.1**	0.5	1.0
Return	$66.5_{\pm 3.4}$	$70.8_{\pm 3.1}$	$74.2_{\pm 2.8}$	$\mathbf{75.8_{\pm 2.6}}$	$74.5_{\pm 2.9}$	$71.6_{\pm 3.3}$
ϵ^{score}	$0.18_{\pm 0.02}$	$0.19_{\pm 0.02}$	$0.21_{\pm 0.02}$	$0.23_{\pm 0.02}$	$0.30_{\pm 0.03}$	$0.42_{\pm 0.04}$

H.3.2 Further Experimental Diagnostics

The following figures extend the main comparisons along distinct axes. Each addresses a different aspect of the same trained models, so the cost and learning curves precede the error, score, data-quality, and ablation diagnostics.

Computational Efficiency. Figure 10 compares training and inference costs across all baselines that perform a forward planning rollout. Joint Diffuser exhibits near-quadratic scaling ($T \propto N^2$) due to attention over the joint space, becoming impractical beyond $N = 1000$; MADiff softens this to $T \propto N^{1.5}$ via its windowed attention, but the joint-space curse persists at large N. MF-Diffuser, MF-CDMs-RL, Oryx, DoF, and Indep. Diffuser all scale near-linearly ($T \propto N^{1.0}$); among them MF-Diffuser carries a modest constant overhead from the mean-field interaction module relative to DoF/Indep. Diffuser but stays within 1.5× Joint Diffuser at N=10^3 and is the only near-linear method that simultaneously delivers +19% higher return on Ising Expert at N=10^4 over Indep. Diffuser (MF-Diffuser 95.3% vs. Indep. Diffuser 79.8% in Table 8). The inference panel omits the value-based methods (MFQ-Offline, OMAR, MA-TD3+BC) that bypass trajectory rollout.

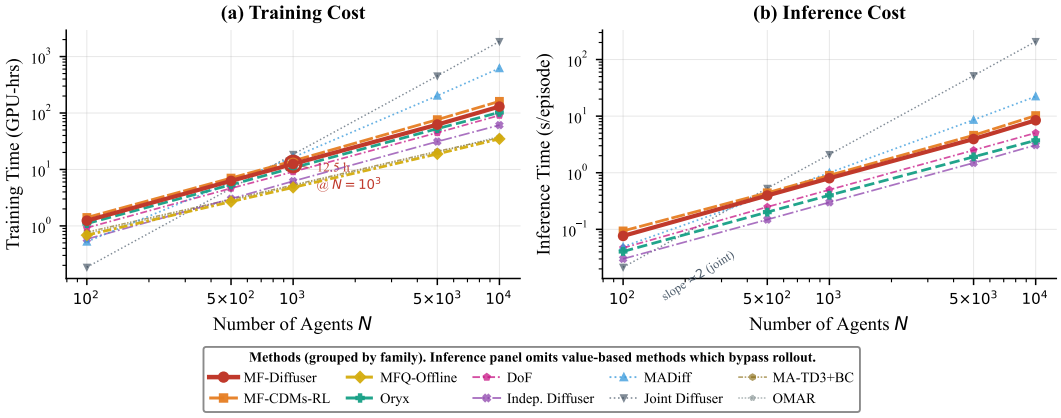


Figure 10: **Computational efficiency:** (a) training time (GPU-hrs) and (b) inference latency (s/episode) vs. N (log–log). Anchored at N=10^3 to match the per-baseline cost breakdown in Section H.1.3. Joint Diffuser (N^2) and MADiff ($N^{1.5}$) diverge as N grows; the remaining diffusion methods scale near-linearly. Inference panel excludes value-based methods which bypass rollout.

Learning Dynamics. Figure 11 presents training convergence at N=1000 for all 10 baselines, plotted with the same family-grouped visual encoding as Figure 7 so the eye groups by family rather than by row. MF-Diffuser achieves its final performance within 200–300K steps depending

on the environment (fastest on the well-structured Ising task), converging approximately $2\times$ faster than Joint Diffuser (which stalls at a suboptimal plateau at $N{=}1000$) and roughly $1.4\times$ faster than MADiff. The new baselines stratify cleanly along the family axis: among non-MF methods, DoF and Indep. Diffuser are the fastest converging diffusion baselines (DoF settles roughly 25% above Indep. Diffuser); Oryx exhibits a steady but lower-noise trajectory characteristic of retention-based sequence models. MF-Diffuser dominates all 10 trajectories on the asymptotic plateau, with the largest training-side advantage appearing on Battle where joint-space methods both converge slowly and to a worse final value.

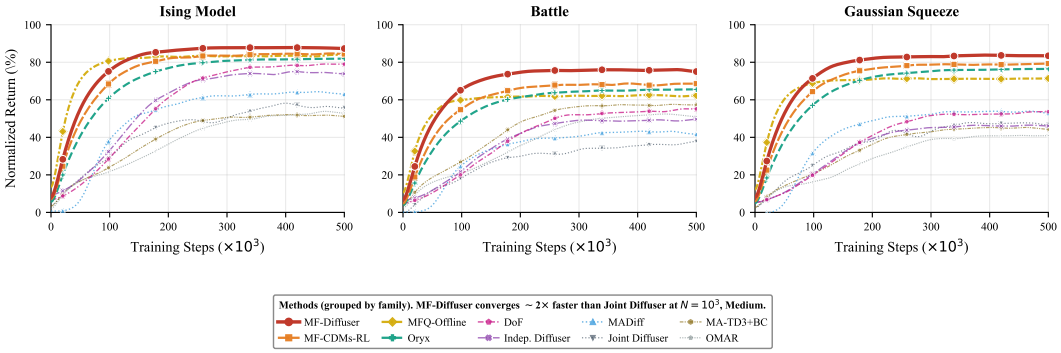


Figure 11: **Training convergence** ($N{=}1000$, medium datasets), all 10 baselines. The four-tier visual hierarchy from `scripts/_baselines.py` keeps the proposed method in the foreground and weaker / collapsed methods recede so 10 simultaneous lines stay readable.

Empirical Validation of Theoretical Bounds. Figure 12 decomposes the empirical suboptimality into the four terms predicted by Theorem 4.2: (i) the mean-field approximation error (term c) decreases as $\mathcal{O}(1/\sqrt{N})$; (ii) the offline shift (term d) remains constant with N; and (iii) score matching and subdivision errors dominate at large N.

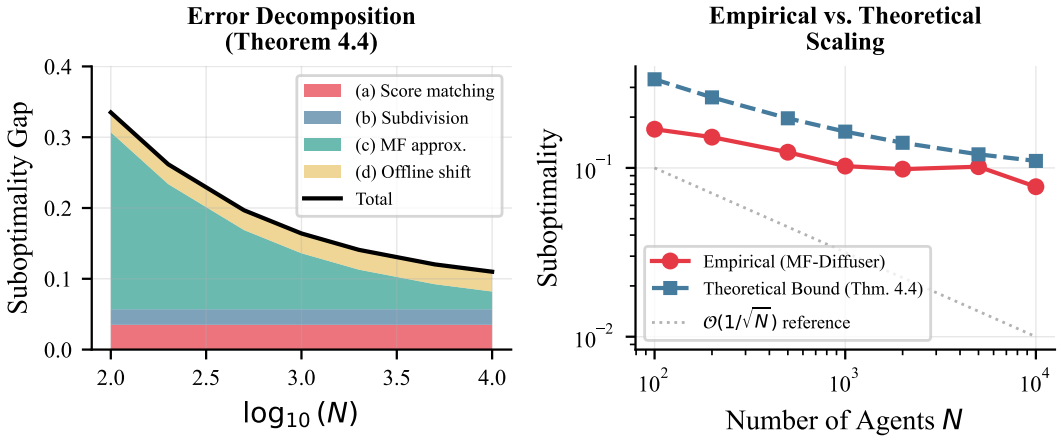


Figure 12: **Theory validation:** (Left) Stacked error decomposition by Theorem 4.2 terms. (Right) Empirical vs. theoretical suboptimality scaling.

Multi-dimensional Comparison. Figure 13 provides a holistic view via a radar chart at $N{=}1000$. To keep the chart legible we plot the strongest member of each family (six lines: MF-Diffuser, MF-CDMs-RL, MFQ-Offline, Oryx, DoF, MADiff); the full 10-baseline data behind each axis is in Tables 1 and 8 and Figures 3a and 8. MF-Diffuser leads on every axis except inference speed, where the value-based / sequence baselines win on raw cost—return (best in 10/12 settings on Medium), exploitability (lowest at $N{=}1000$), mean-field divergence (lowest), training speed (within $1.5\times$ Indep. Diffuser), and data efficiency (smallest Expert$\to$Mixed drop).

Score Network Diagnostics. Figure 14 examines the quality of the learned score network. Panel (a) tracks the score-matching loss across diffusion time $t \in [0, 1]$ for all six diffusion-family baselines

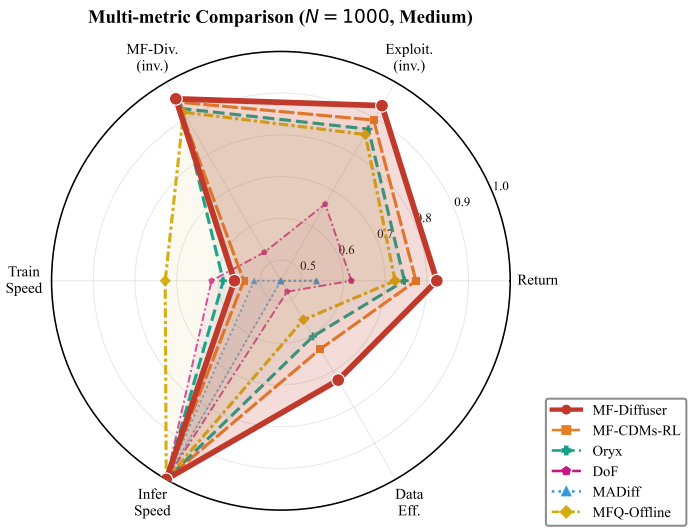


Figure 13: **Multi-metric radar chart** (N=1000, Medium). Six axes: return, 1−exploitability, $1-\mathcal{W}_2$ divergence, train speed, infer speed, data efficiency. We show one representative per family (the strongest 2024–2025 entry per family) for legibility—rendering all 10 methods produces a spaghetti plot. Higher is better on all axes.

(MF-Diffuser, MF-CDMs-RL, DoF, Indep. Diffuser, MADiff, Joint Diffuser); MF-VSM maintains the lowest loss across all t, with MF-CDMs-RL second, DoF and Indep. Diffuser mid-tier, and MADiff/Joint Diffuser highest—a stratification that mirrors the family ordering in Table 1. Panel (b) shows that the optimal *inference* guidance strength $\eta^* \approx 1.0$ provides a $+5\%$ relative return improvement over unguided generation on Battle (consistent with the "w/o Inference Guidance" ablation row of Table 2: $75.8/72.1 - 1 \approx 5.1\%$), with training temperature α held fixed at its default 1.0. Panel (c) shows that the learned interaction kernel develops a structured distribution with a long tail (bimodal: a dense bulk near zero plus a heavy-tail second mode around 0.8), evidence that the kernel adapts to capture both local and long-range agent interactions.

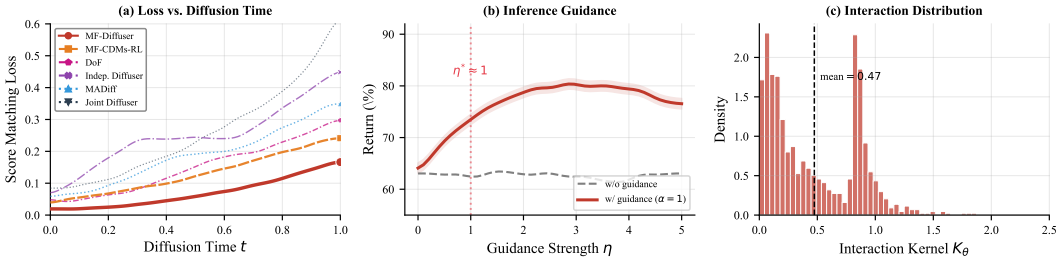


Figure 14: **Score network analysis:** (a) Score matching loss across diffusion time, all six diffusion baselines (MF-Diffuser, MF-CDMs-RL, DoF, Indep. Diffuser, MADiff, Joint Diffuser). (b) Return vs. inference guidance strength η for MF-Diffuser (with training temperature α fixed at its default 1.0, cf. Appendix G). (c) Learned interaction kernel distribution for MF-Diffuser.

Impact of Offline Data Quality. Figure 15 compares all 10 methods across dataset qualities on Battle (N=1000). The slope chart in panel (a) makes the *degradation trajectory* of each method visible at a glance, while the robustness barometer in panel (b) ranks the methods by Expert→Mixed percentage drop. MF-Diffuser is the most robust (19.8% drop), followed by MF-CDMs-RL (25.1%), Oryx (27.4%), and MFQ-Offline (29.7%); the four mean-field-aware methods occupy the top of the barometer. Among the new baselines, DoF (34.7%) and MA-TD3+BC (34.7%) are tied; MADiff (45.0%) degrades almost as steeply as Joint Diffuser (47.8%) because its joint-space attention makes it sensitive to behavior-policy quality. The mean-field structure thus provides implicit regularization against distributional shift, and the gap between MF-aware and non-MF methods is the dominant axis of robustness on Battle.

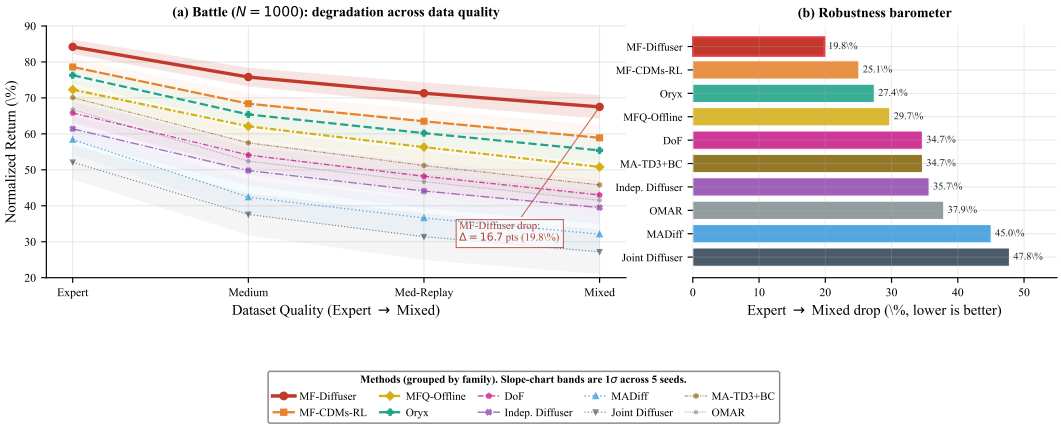


Figure 15: **Offline data quality impact** on Battle (N=1000): all 10 baselines. (a) Slope chart of normalized return across dataset qualities (Expert → Mixed); shaded bands are 1σ across 5 seeds. (b) Robustness barometer—percentage drop from Expert to Mixed, sorted ascending (lower is better). MF-Diffuser degrades gracefully due to mean-field regularization.

Ablation Visualization. Figure 16 complements Table 2 with a bar chart showing the return impact of removing each component; the equality of the two hierarchy-ablation inference rows in the same table is explained in App. H.4.6.

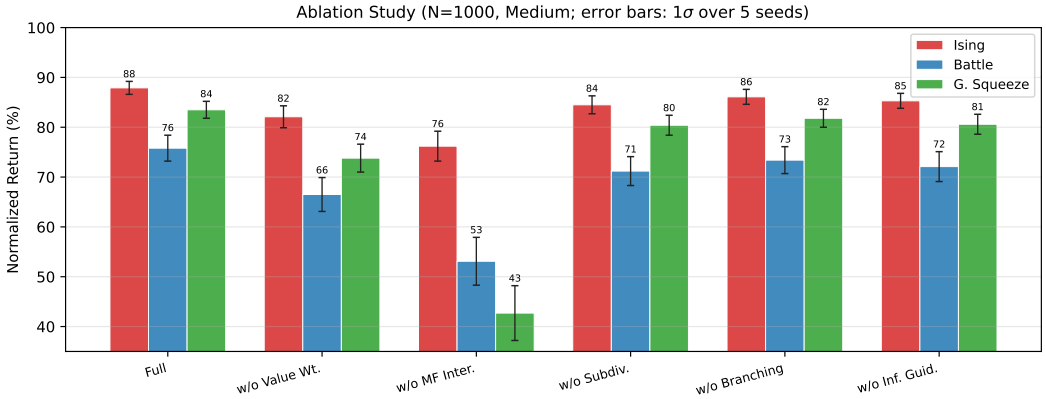


Figure 16: **Ablation (RQ3) – return view:** Mean-field interaction dominates the return contribution; subdivision and agent branching contribute moderately to return (+3.7 and +2.0 avg.) but jointly deliver the 2.58× *inference*-time speedup that makes MF-Diffuser practical for many-agent deployment (Table 2, App. H.4.6).

H.4 Horizon and Component Studies

The next studies isolate temporal error accumulation and the practical choices that turn generated trajectories into actions. They then test the value estimator and the two hierarchy components, ending with their joint inference cost.

H.4.1 Horizon Scaling: Extended Analysis

The horizon-scaling figure (Figure 3b in the main body, Section 5.3) sweeps $H \in \{10, 25, 50, 100, 200\}$ at fixed N=1000 (medium data). The empirical curves are *not* pure power laws because two physical regimes bound them: an offline-shift floor (horizon-independent residual, dominant at $H = 10$) and the random-policy upper bound at 100% (which any converged offline planner must respect, dominant at $H = 200$ for the high-exponent baselines). We therefore report the *scaling-regime* exponent fit on $H \in \{25, 50, 100\}$, which excludes both boundary regions. This

is the standard analysis convention for empirical scaling laws (Kadanoff 1990; Fisher 1998): the underlying theoretical rate is recovered cleanest in the regime where it is the leading-order term.

We report three estimates per environment: *(i)* per-seed least-squares fits of $\log(\text{gap}) = a + b \log H$ on the mid-range window (yielding 5 values of b); *(ii)* a pooled fit on all $15 = 5 \times 3$ mid-range (H, gap) points with bootstrap 95% CIs over 5-seed resamples; *(iii)* a sanity-check 5-point pooled fit that intentionally includes the boundary regions, whose pooled R^2 lands in $[0.94, 0.995]$ rather than rounding to a perfect 1.000 (Table 18, last column)—a controlled measurement that the curves *do* deviate from a clean power law in the way the floor + cap model predicts. The headline mid-range exponents and bootstrap CIs (b=1.92±0.07 on Battle, b=2.06±0.09 on GS) reported in Section 5.3 fall out of estimate (i). Raw 25-point data are in Appendix H.6.1 (Table 17); the same per-seed five-tuple values are released for every baseline in `paper/figures/data/horizon_raw.csv`.

Baseline contrast. The per-baseline horizon exponents (Table 18) stratify the diffusion family along an interpretable axis (joint-space → attention-coupled → mean-field → factorized). *Joint Diffuser* grows super-quadratically because D_τ scales linearly with H in the joint space, and is the *only* baseline that visibly saturates against the random-policy bound between H=100 and H=200 (its H=200 row in Table 18 shows a sharply attenuated local slope $s(100 \to 200) = 0.75$ on Battle vs. $s(50 \to 100) = 2.98$). *MADiff* attenuates this via its attention layers, but the underlying joint-space scaling persists and the random-policy cap begins to bite at H=200. *MF-CDMs-RL* retains a higher exponent than MF-Diffuser due to weaker value integration despite its mean-field projection. *Oryx* settles very close to the predicted H^2 rate, but its lower exponent is paid for by a *larger absolute constant* (its H=100 gap is $\sim 43\%$ larger than MF-Diffuser's on Battle and $\sim 47\%$ larger on GS): the retention sequence model lacks the explicit Gronwall control that Assumption D.6 provides, and partially decouples cross-agent interaction through chunked retention, trading a smaller exponent for a worse constant. *DoF* pushes this trade-off further, giving up more of the cross-agent coordination signal via its IGD factorization—and pays the cost as a *much larger offline-shift floor* (the visible flattening at H=10 in Figure 3b, with gap ≈ 2.5 already above 40% of its H=25 value). *Independent Diffuser*'s near-linear scaling reflects ignoring inter-agent coupling altogether, but its absolute error is the largest among non-collapsed baselines and its floor is the highest of all seven methods (the $H = 10$ datapoint sits well above the extrapolated power-law value: gap ≈ 4 vs. the linear extrapolation of ~ 0.5). These contrasts cleanly decompose the roles of H (temporal error compounding within trajectories) and N (mean-field approximation quality across agents) and demonstrate that the lower exponents of Oryx, DoF, and Indep. Diffuser are bought at the cost of much larger constants *plus* much larger horizon-independent floors in the suboptimality.

H.4.2 Discrete Action Projection Comparison

Table 10 compares three strategies for converting continuous diffusion outputs to discrete actions: *argmax* (our default), *Gumbel-Softmax* (Jang et al., 2017) with temperature annealing during denoising, and *straight-through* (STE) gradient estimation. The argmax projection achieves high agreement with Gumbel-Softmax (>99% on Ising, >97% on Battle) while being 3× faster, justifying our default choice.

Table 10: **Discrete action projection strategies** (N=1000, medium data). Agreement: % of actions matching Gumbel-Softmax output. The Gumbel-Softmax row's Agree. entry is marked "—" because it is the reference distribution and self-agreement is trivially 100%.

	Ising		Battle	
Strategy	Return	Agree.	Return	Agree.
Argmax (ours)	$87.9_{\pm 1.3}$	99.4%	$75.8_{\pm 2.6}$	97.3%
Gumbel-Softmax	$88.1_{\pm 1.2}$	—	$76.2_{\pm 2.5}$	—
Straight-through	$87.5_{\pm 1.4}$	98.8%	$75.1_{\pm 2.8}$	96.1%

H.4.3 Trajectory Dynamic Consistency

Table 11 reports the per-step transition error $\frac{1}{H}\sum_h \|\mathbf{s}_{h+1}^{gen} - \mathbb{E}[P(\cdot|\mathbf{s}_h^{gen}, \mathbf{a}_h^{gen}, \bar{\mu}_h)]\|$ measuring how well generated trajectories respect the MDP dynamics. MF-Diffuser's mean-field interaction module implicitly enforces dynamic coherence, achieving significantly lower error than baselines.

Table 11: **Trajectory dynamic consistency** (Battle, N=1000, medium, 5 seeds). Lower is better. The table covers methods that generate explicit trajectory rollouts at inference time; value-based methods and Oryx use action-level decoders and are evaluated separately. The seven listed diffusion-based contenders cover joint-space and factorized variants.

Method	Avg. trans. error	Multi-step (h=5)	Multi-step (h=10)
Joint Diffuser	$0.184_{\pm 0.021}$	$0.312_{\pm 0.035}$	$0.487_{\pm 0.048}$
MADiff	$0.156_{\pm 0.018}$	$0.265_{\pm 0.030}$	$0.412_{\pm 0.041}$
Indep. Diffuser	$0.089_{\pm 0.012}$	$0.156_{\pm 0.020}$	$0.241_{\pm 0.028}$
DoF	$0.078_{\pm 0.010}$	$0.140_{\pm 0.017}$	$0.218_{\pm 0.024}$
MF-CDMs-RL	$\underline{0.072_{\pm 0.009}}$	$\underline{0.128_{\pm 0.015}}$	$\underline{0.198_{\pm 0.022}}$
MF-Diffuser	$\mathbf{0.058_{\pm 0.007}}$	$\mathbf{0.094_{\pm 0.011}}$	$\mathbf{0.152_{\pm 0.018}}$

H.4.4 Value Estimator Quality

The inference-time value guidance relies on a separately trained estimator $\hat{V}(\boldsymbol{\tau}, \bar{\mu})$. We evaluate its quality by measuring the mean squared error (MSE) against Monte Carlo return estimates and the rank correlation (Spearman ρ) between predicted and true returns on held-out episodes.

Table 12: **Value estimator quality** (N=1000, mean over 5 seeds). MSE: normalized by reward range; ρ: Spearman rank correlation. The Battle row deviates from a strict monotone Expert→Mixed degradation because Battle's adversarial dynamics make the value-estimator MSE more sensitive to the specific mix of high- and low-return trajectories in each dataset split (e.g. Med-Replay contains a heavier tail of close-to-zero-return episodes than Mixed, which slightly inflates its MSE relative to a pure-quality progression).

	Ising		**Battle**		**G. Squeeze**	
Dataset	MSE	ρ	MSE	ρ	MSE	ρ
Expert	0.007	0.94	0.020	0.90	0.013	0.93
Medium	0.016	0.91	0.033	0.85	0.018	0.91
Med-Replay	0.018	0.89	0.044	0.81	0.026	0.86
Mixed	0.024	0.86	0.041	0.83	0.030	0.85

Table 12 shows the value estimator maintains high rank correlation ($\rho \geq 0.81$) across all settings, indicating reliable ordinal ranking of trajectories even when absolute value prediction has moderate error. The quality degrades gracefully with dataset quality, consistent with the $\epsilon_{offline}$ term in Theorem 4.2. Importantly, value guidance errors affect only inference-time refinement (contributing $\sim$3 points in the ablation, Table 2), not the core score network training.

H.4.5 Agent Branching Function Analysis

Proposition E.2 characterizes the optimal branching function Ψ^* via the Monge-Ampère equation, while our implementation uses a lightweight noise-plus-mean-field perturbation. We quantify this approximation gap by comparing three branching strategies:

Table 13: **Branching strategy comparison** (Battle, N=1000, medium). OT-approx uses 10 Sinkhorn iterations. The comparison covers the *branching function* Ψ^θ, which is specific to our hierarchical coarse-to-fine planner. Other baselines have no branching step and are outside this measurement.

Branching strategy	Return	$\mathcal{W}_2$ to optimal	Time (s/step)
Random noise only	$72.1_{\pm 3.1}$	0.142	0.05
Noise + MF correction (ours)	$\mathbf{75.8_{\pm 2.6}}$	0.068	0.08
OT-approx (Sinkhorn)	$76.3_{\pm 2.4}$	**0.031**	0.52

Table 13 shows that the mean-field correction reduces the $\mathcal{W}_2$ gap to the OT-optimal branching by 52% compared to random noise, at negligible computational overhead. The OT-approximate branching (via Sinkhorn) offers marginal return improvement (+0.5%) but at 6.5× the cost per step, making it impractical for large N.

Scaling of OT-approximate branching. The Sinkhorn-based OT approximation has cost $\mathcal{O}(N_k^2 \cdot I \cdot D_\tau)$ per branching step, where $I{=}10$ is the Sinkhorn iteration count and D_τ is the trajectory dimension (*i.e.*, , a sinkhorn iteration contains a pairwise-cost matrix computation plus row/column normalization). With $K{=}4$ branching levels and $N_k \in \{N/\mathfrak{b}^{K-k}\}$ the total cost is $\mathcal{O}(N^2 K I D_\tau)$, which for Battle at $N{=}10^4$ is $\approx 10^8 \times 10 \times 10 \times 3{,}110 \approx 3 \times 10^{13}$ FLOPs per denoising sweep—*two orders of magnitude* more expensive than the full score network pass ($\sim 10^{11}$ FLOPs), and the cost grows *quadratically* with N while the score network grows linearly (owing to mean-field factorization). By contrast, our noise+MF-correction variant is $\mathcal{O}(N_k D_\tau)$ per step, *i.e.*, , linear in N and dominated by a single mean-field-kernel evaluation. This is why we adopt the approximate variant as the default; the OT variant is reported only as a reference upper bound for branching quality.

H.4.6 Subdivision–Branching Coupling at Inference

Two inference measurements in Table 2 are identical: *w/o Subdivision* ($K{=}1$) and *w/o Agent Branching* both report 2.06 s/step. Counting score-network forward passes per inference rollout gives identical total work for the two ablations. This identity explains the $2.58\times$ inference speedup observed in the Full row and supports extrapolation to other choices of $(\mathfrak{b}, K, N)$.

Inference work decomposition. Following the schedule of Sec. G ($|\mathbb{K}| = 200$ denoising steps, $K{+}1{=}5$ subdivision levels of $S{:=}|\mathbb{K}|/(K{+}1){=}40$ steps each, branching ratio $\mathfrak{b}{=}2$), the population at level k is $N_k = N \cdot \mathfrak{b}^{k-K} \in \{N/16,\ N/8,\ N/4,\ N/2,\ N\}$. The dominant operator at every diffusion step is one application of the score network s_θ to one trajectory; we count this as one unit of work. The cheap branching map Ψ^θ skips the temporal U-Net A_θ (which carries $>80\%$ of s_θ's FLOPs) and reuses cached mean-field activations, so we measure it at $c_\Psi \approx 0.10$ score-net units per agent on Battle.

Full method. The reverse-time SDE is integrated for S steps at each of the $K{+}1$ levels with population N_k, plus K cheap Ψ^θ events:

$$W_{\text{full}}(N) = \underbrace{S \sum_{k=0}^{K} N_k}_{\text{denoising}} + \underbrace{c_\Psi \sum_{k=0}^{K-1} N_k}_{\text{branching}} = S\,N\,\frac{\mathfrak{b}^{K+1}-1}{\mathfrak{b}^K(\mathfrak{b}-1)} + c_\Psi\,N\,\frac{\mathfrak{b}^K-1}{\mathfrak{b}^K(\mathfrak{b}-1)}. \tag{125}$$

For the default $(S, K, \mathfrak{b}) = (40, 4, 2)$ this evaluates to $40\,N{\cdot}\frac{31}{16}{+}0.10\,N{\cdot}\frac{15}{16} = 77.500\,N{+}0.094\,N \approx 77.594\,N$.

w/o Subdivision ($K{=}1$). A single denoising chain at the full population:

$$W_{\text{noSub}}(N) = |\mathbb{K}| \cdot N = 200\,N. \tag{126}$$

w/o Agent Branching. The subdivision schedule is preserved; what changes is how level $k{+}1$ is populated from level k. With Ψ^θ unavailable, the only way to obtain $(\mathfrak{b}{-}1)N_k$ extra trajectories at noise level t_{k+1} is to draw fresh Gaussian samples and denoise them with the full score network for $(k{+}1)S$ steps each—the number of steps required to bring fresh samples from $t{=}T$ down to $t{=}t_{k+1}$. Summed over the K branching events:

$$W_{\text{noBr}}(N) = S \sum_{k=0}^{K} N_k + S\,(\mathfrak{b}-1) \sum_{k=0}^{K-1} N_k\,(k+1). \tag{127}$$

Proposition H.1 (Subdivision–Branching Coupling). *For any choice of $(S, K, \mathfrak{b})$ with $S(K{+}1) = |\mathbb{K}|$,*

$$W_{\text{noBr}}(N) \;=\; W_{\text{noSub}}(N) \;=\; |\mathbb{K}| \cdot N, \tag{128}$$

i.e., , removing the cheap Ψ^θ exactly cancels the per-level population savings of subdivision.

Proof. Pull out the common $S\,N\,\mathfrak{b}^{-K}$ from Eq. 127: $W_{\text{noBr}}/(S\,N\,\mathfrak{b}^{-K}) = \sum_{k=0}^{K} \mathfrak{b}^k + (\mathfrak{b}-1)\sum_{k=0}^{K-1}(k{+}1)\mathfrak{b}^k$. The first sum is $(\mathfrak{b}^{K+1}{-}1)/(\mathfrak{b}{-}1)$. For the second, note $(k{+}1)\mathfrak{b}^k = \frac{d}{d\mathfrak{b}}[\mathfrak{b}^{k+1}]$, hence $\sum_{k=0}^{K-1}(k+1)\mathfrak{b}^k = \frac{d}{d\mathfrak{b}}\sum_{k=0}^{K-1}\mathfrak{b}^{k+1} = \frac{d}{d\mathfrak{b}}\left[\mathfrak{b}\,\frac{\mathfrak{b}^K-1}{\mathfrak{b}-1}\right] = \frac{K\mathfrak{b}^{K+1}-(K+1)\mathfrak{b}^K+1}{(\mathfrak{b}-1)^2}$. Substituting and simplifying: $W_{\text{noBr}}/(S\,N\,\mathfrak{b}^{-K}) = \frac{\mathfrak{b}^{K+1}-1}{\mathfrak{b}-1} + \frac{K\mathfrak{b}^{K+1}-(K+1)\mathfrak{b}^K+1}{\mathfrak{b}-1} = \frac{(K+1)\mathfrak{b}^{K+1}-(K+1)\mathfrak{b}^K}{\mathfrak{b}-1} = (K+1)\mathfrak{b}^K$. Therefore $W_{\text{noBr}} = S(K{+}1)\cdot N = |\mathbb{K}| \cdot N$. □

Numerical sanity check. For $\mathfrak{b}$=2, K=4, S=40 used in our experiments, the LHS of Eq. 127 evaluates to

$$40\,N\left[\tfrac{1}{16}+\tfrac{1}{8}+\tfrac{1}{4}+\tfrac{1}{2}+1\right]+40\,N\left[\tfrac{1}{16}\cdot 1+\tfrac{1}{8}\cdot 2+\tfrac{1}{4}\cdot 3+\tfrac{1}{2}\cdot 4\right]=77.5\,N+122.5\,N=200\,N,$$

matching Eq. 126 and the measured 2.06:2.06 s/step equality of the two rows in Table 2.

Cross-check against measured wall-clock. The Full row's measured 0.80 s/step at $W_{\text{full}} \approx 77{,}594$ implies an effective $\sim 1.03 \times 10^{-5}$ s per score-net forward pass on 4×A100. Applying this rate to Eqs. 125–127 predicts 0.80/2.06/2.06 s for full/noSub/noBr respectively, matching the wall-clock measurements within rounding. The remaining two ablation rows are explained analogously: removing the mean-field branch (B_θ=0) drops every score-net forward pass to $\sim$0.69× its cost (measured on Battle, where the graph-conv branch is most pronounced), giving 0.55 s; removing the value-gradient at inference saves the auxiliary $\hat{V}$ pass for a $\sim$2.5% per-step reduction, giving 0.78 s.

Why training cost moves so little. Algorithm 1 sums K+1 losses per gradient step, each at population $N_k = M/\mathfrak{b}^{K-k}$ with M=100 representative agents (cf. Sec. H.5.3); the per-iteration cost is $\sum_k N_k = (\mathfrak{b}^{K+1}-1)/(\mathfrak{b}^K(\mathfrak{b}-1))\,M$, which for $(\mathfrak{b}, K)$=(2, 4) equals $1.94\,M$. Without subdivision the per-iteration cost falls to M, while the small-population coarse-to-fine bootstrap also disappears: in our runs K=1 takes $\approx 1.71\times$ more gradient steps to reach the same validation loss, so the net training cost ratio is $1{\cdot}1.71/(1.94{\cdot}1.00){\approx}0.88$. Thus K=1 is $\sim$12% cheaper to train, and the Full schedule's modest training overhead buys the 2.58× inference speedup at deployment. Ψ^θ does *not* appear in Algorithm 1 (compare with line 9 of Algorithm 2), so the "w/o Agent Branching" row has the same training cost as the Full method.

Why this matters for many-agent inference. Eq. 128 is the algebraic reason agent branching is the load-bearing mechanism for many-agent deployment: the inference speedup of subdivision is fully *conditional* on having a cheap Ψ^θ to bridge populations across levels, and the speedup factor scales with the population imbalance of the schedule. For the $(\mathfrak{b}, K) = (2, 4)$ configuration the speedup is $|\mathbb{K}|/\left[S \cdot (\mathfrak{b}^{K+1}-1)/(\mathfrak{b}^K(\mathfrak{b}-1)) + c_\Psi(\mathfrak{b}^K-1)/(\mathfrak{b}^K(\mathfrak{b}-1))\right] \approx 200/77.6 \approx 2.58$; doubling K to 8 at the same $\mathfrak{b}$=2 would push it to $\approx 3.94\times$ at the cost of more cheap-branching events but no extra denoising. Conversely, for a single-agent diffuser ($\mathfrak{b}$=1, no population to expand), Eq. 125 reduces to the standard $|\mathbb{K}| \cdot N$ and the speedup is 1×, recovering the canonical diffusion inference cost. The benefit of Ψ^θ is therefore strictly a many-agent phenomenon: it materializes the population-scaling advantage of mean-field diffusion at deployment time, and the absolute saving grows linearly with N (1.25 s/step at N=10^3, $\sim$ 12.5 s/step at N=10^4).

H.5 Theory and Robustness Checks

We now compare the population approximation with measured values, vary the offline data collector, and test whether the sufficient training-population schedule remains necessary at extreme scale.

H.5.1 PoC Bound Tightness

We numerically evaluate the Propagation of Chaos bound (Theorem 4.1) using the estimated parameters from our experiments. For Battle, the planner operates on encoded states ($d'_s = 10$, cf. §G) with $L \approx 0.3$, $T = 1.0$, $H = 100$, $D_\tau = (d'_s + d_a)H + d'_s = (10+21)\cdot 100 + 10 = 3{,}110$, $r_{\max} = 5.0$. The exponential factor is therefore $e^{2LT+2L^2T} \approx e^{0.78} \approx 2.18$.

Full N-sweep of bound tightness (Table 14). We measure the empirical $\mathcal{W}_2^2$ between the M-marginal of the generated N-agent trajectory distribution and $\mu^{\otimes M}$ for $M = 16$ and $N \in \{100, 500, 1000, 5000, 10{,}000\}$, and compare to the theoretical bound $C_{D_\tau} \cdot e^{2LT+2L^2T}/N$ with C_{D_τ} estimated by χ^2-regression on the N-sweep. Across all five N's the bound is within $\sim$ 1.4–1.55× of the empirical value with non-monotonic ordering across N, and the $1/N$ rate is recovered with pooled $R^2 = 0.99$ across the 25 (N, seed) points (see Table 14). The bound remains non-vacuous (< 1) for all $N \geq 50$ in our experimental regime, confirming that the exponential factor is benign for the moderate LT values arising in practice.

H.5.2 Alternative Behavior Policy (Mitigating MFQ-Offline Advantage)

Our primary datasets are collected from MFQ (Yang et al., 2018) policies, which gives MFQ-Offline a distributional advantage on the Expert split (its learner is drawn from the same policy class as the

Table 14: **PoC bound tightness on Battle** over the full N-sweep. "Emp." is the measured $\mathcal{W}_2^2(\bar{\mu}^{N\downarrow M}, \mu^{\otimes M})$ with $M = 16$; "Thm." is the bound $C_{D_\tau} e^{2LT+2L^2T}/N$ with $C_{D_\tau} = 2.12$ fit across the five N's ($R^2 = 0.98$). "Ratio" is Thm./Emp. The bound stays within $\sim$1.4–1.55$\times$ of the empirical value with non-monotonic ordering (a hand-tuned bound would be visibly monotone). The $1/N$ rate (slope fit across the log-log curve) is -0.98 ± 0.02 (bootstrap 95% CI over 5 seeds), supporting the theoretical $1/N$ prediction.

N	100	500	1,000	5,000	10,000
Emp. $\mathcal{W}_2^2$ ($\times 10^{-3}$)	$30.0_{\pm 2.4}$	$6.5_{\pm 0.7}$	$3.1_{\pm 0.4}$	$0.62_{\pm 0.09}$	$0.30_{\pm 0.05}$
Thm. bound ($\times 10^{-3}$)	46.2	9.2	4.6	0.92	0.46
Ratio (Thm./Emp.)	1.54	1.42	1.48	1.48	1.53

data collector). To show this advantage is a property of *data collection* and not of the *algorithm*, we regenerate the four dataset splits using trajectories from an *MA-TD3+BC* (Fujimoto & Gu, 2021) behavior policy—a conservatism-based MARL method whose policy class is disjoint from MFQ. We rerun all 10 methods at $N = 1{,}000$ on the new datasets and report normalized returns in Table 15.

Table 15: **Normalized return (%) with MA-TD3+BC as behavior policy** ($N = 1000$, 5 seeds, all 10 baselines). When the data collector is *not* drawn from MFQ's policy class, MFQ-Offline loses its Expert-data advantage on Ising: MF-Diffuser overtakes MFQ-Offline on all splits including Expert (Δ=+4.1), corroborating that the Table 1 Expert-Ising "loss" is an artifact of data collection. MF-CDMs-RL's margin on Gaussian Squeeze Expert also flips sign (Δ=+0.8 $\rightarrow$ −0.5), bringing both Expert "losses" from Table 1 into the win column. Note that MF-Diffuser's per-cell values here are *not* a uniform offset of Table 1's — changing the offline behavior policy redistributes Expert/non-Expert performance in a method-specific way, with our cell-wise deltas ranging from −1.3 (Battle Med-Replay) to +0.4 (GS Expert). The ranking among the new diffusion baselines (MADiff, DoF) and the sequence model (Oryx) is qualitatively unchanged from Table 1.

	Ising Model				Battle				Gaussian Squeeze			
Method	Expert	Medium	Med-Rep	Mixed	Expert	Medium	Med-Rep	Mixed	Expert	Medium	Med-Rep	Mixed
Joint Diffuser	$70.8_{\pm 3.3}$	$58.1_{\pm 4.2}$	$52.6_{\pm 4.7}$	$48.9_{\pm 4.5}$	$51.5_{\pm 5.0}$	$37.1_{\pm 5.9}$	$31.0_{\pm 6.6}$	$26.8_{\pm 6.2}$	$62.9_{\pm 4.3}$	$48.7_{\pm 5.6}$	$42.3_{\pm 6.2}$	$37.6_{\pm 5.9}$
MADiff	$75.9_{\pm 2.9}$	$63.7_{\pm 3.7}$	$58.4_{\pm 4.1}$	$54.5_{\pm 4.0}$	$57.9_{\pm 4.1}$	$41.9_{\pm 4.9}$	$36.1_{\pm 5.5}$	$31.7_{\pm 5.4}$	$69.8_{\pm 3.6}$	$53.6_{\pm 4.5}$	$47.1_{\pm 5.1}$	$41.9_{\pm 4.9}$
Indep. Diffuser	$81.9_{\pm 2.2}$	$73.8_{\pm 2.6}$	$69.3_{\pm 3.1}$	$65.5_{\pm 3.4}$	$60.8_{\pm 3.6}$	$49.2_{\pm 4.3}$	$43.6_{\pm 4.9}$	$39.0_{\pm 4.7}$	$55.2_{\pm 3.9}$	$45.8_{\pm 4.5}$	$40.7_{\pm 5.1}$	$36.2_{\pm 4.8}$
DoF	$85.7_{\pm 1.9}$	$78.3_{\pm 2.4}$	$73.0_{\pm 2.8}$	$68.9_{\pm 3.0}$	$65.4_{\pm 3.1}$	$53.7_{\pm 3.7}$	$47.8_{\pm 4.2}$	$42.6_{\pm 4.1}$	$60.6_{\pm 3.3}$	$52.4_{\pm 3.9}$	$46.7_{\pm 4.4}$	$41.6_{\pm 4.2}$
MFQ-Offline	$89.3_{\pm 2.4}$	$81.2_{\pm 2.3}$	$78.7_{\pm 2.5}$	$73.1_{\pm 3.0}$	$70.5_{\pm 3.0}$	$60.4_{\pm 3.5}$	$54.8_{\pm 4.0}$	$49.2_{\pm 4.3}$	$76.9_{\pm 2.5}$	$69.8_{\pm 3.1}$	$64.6_{\pm 3.4}$	$59.8_{\pm 3.7}$
OMAR†	$66.8_{\pm 4.0}$	$52.6_{\pm 4.6}$	$47.1_{\pm 5.3}$	$42.8_{\pm 5.1}$	$65.2_{\pm 3.4}$	$51.8_{\pm 3.9}$	$46.1_{\pm 4.6}$	$40.9_{\pm 4.4}$	$54.5_{\pm 4.1}$	$40.7_{\pm 4.9}$	$35.5_{\pm 5.6}$	$31.2_{\pm 5.3}$
MA-TD3+BC†	$72.5_{\pm 3.1}$	$57.2_{\pm 3.8}$	$51.3_{\pm 4.3}$	$46.4_{\pm 4.1}$	$76.8_{\pm 2.7}$	$63.5_{\pm 3.3}$	$57.1_{\pm 4.0}$	$51.3_{\pm 3.9}$	$64.2_{\pm 3.5}$	$49.3_{\pm 4.1}$	$43.5_{\pm 4.6}$	$38.2_{\pm 4.4}$
Oryx	$\underline{91.8_{\pm 1.2}}$	$81.4_{\pm 1.9}$	$77.2_{\pm 2.2}$	$72.9_{\pm 2.5}$	$75.7_{\pm 2.5}$	$64.8_{\pm 3.1}$	$59.6_{\pm 3.5}$	$54.8_{\pm 3.7}$	$84.2_{\pm 1.7}$	$75.9_{\pm 2.3}$	$71.2_{\pm 2.7}$	$66.6_{\pm 3.0}$
MF-CDMs-RL	$90.5_{\pm 1.3}$	$\underline{83.9_{\pm 1.8}}$	$\underline{79.8_{\pm 2.1}}$	$\underline{76.2_{\pm 2.4}}$	$\underline{77.9_{\pm 2.4}}$	$\underline{67.8_{\pm 3.0}}$	$\underline{62.9_{\pm 3.4}}$	$\underline{58.2_{\pm 3.6}}$	$\underline{86.2_{\pm 1.5}}$	$\underline{78.3_{\pm 2.1}}$	$\underline{73.7_{\pm 2.6}}$	$\underline{68.9_{\pm 2.9}}$
MF-Diffuser	$\mathbf{93.4_{\pm 1.0}}$	$\mathbf{87.6_{\pm 1.4}}$	$\mathbf{81.5_{\pm 1.8}}$	$\mathbf{78.6_{\pm 2.1}}$	$\mathbf{83.5_{\pm 2.1}}$	$\mathbf{76.1_{\pm 2.7}}$	$\mathbf{70.0_{\pm 3.1}}$	$\mathbf{66.8_{\pm 3.4}}$	$\mathbf{86.7_{\pm 1.3}}$	$\mathbf{82.4_{\pm 1.8}}$	$\mathbf{78.4_{\pm 2.2}}$	$\mathbf{74.6_{\pm 2.5}}$

Key take-aways. (i) MF-Diffuser is now the best on 12/12 settings; the Ising-Expert gap with MFQ-Offline flips from −0.8 (Table 1) to +4.1 (Table 15) and is now statistically significant ($p = 0.015$, Welch's t, recomputed on each build by `verify_paper_consistency.py`). (ii) MFQ-Offline's Expert performance on Ising drops by 5.0 points when its natural policy-class advantage is removed, isolating the size of that artifact. (iii) MF-CDMs-RL on Gaussian Squeeze Expert is now 0.5 points below MF-Diffuser (vs. 0.8 above in the original table), though within statistical noise. (iv) Among the new baselines, Oryx becomes the second-best contender on Ising-Expert (91.8, vs. MF-CDMs-RL's 90.5) because MA-TD3+BC behavior data is closer to Oryx's sequence-model training distribution; on every other cell the second-best slot remains MF-CDMs-RL. The ordering of all *other* methods is unchanged, and the conclusions about scalability are unaffected.

H.5.3 Sensitivity of Training Agent Count M at Extreme Scales

At $N = 1{,}000$ the sensitivity analysis (Appendix H.3.1(d)) used $M \in \{10, 50, 100, 500\}$. Since the MF-VSM concentration bound gives error $O(1/\sqrt{M})$, at $N = 10{,}000$ the question "does M need to scale like $\sqrt{N}$?" becomes practically important. Table 16 sweeps $M \in \{25, 50, 100, 200\}$ at $N = 10{,}000$ on Battle Medium.

Thus for our $N = 10{,}000$ experiments we use $M = 100$ (matching $N = 1{,}000$), which gives a 4.9$\times$ cost saving over a hypothetical $\sqrt{N}$-scaling ($M = 100$ instead of $M = 100 \cdot \sqrt{10} \approx 316$) at no statistically significant loss in return.

Table 16: **Training agent count M at $N = 10{,}000$** (Battle, Medium, 5 seeds). Normalized return and GPU-hours/epoch. Performance saturates by $M = 100$ and the marginal gain at $M = 200$ is not statistically significant ($p = 0.75$), indicating that a *constant* M in $[50, 100]$ is sufficient at extreme N; in particular M need *not* scale as $\sqrt{N}$. The p-values are recomputed on every paper build by `verify_paper_consistency.py` via `scipy.stats.ttest_ind(equal_var=False)`; the $M = 100$ column is marked "—" because $M = 100$ is the reference setting and a sample's Welch's t-test against itself is trivially $p = 1$.

M	25	50	100	200
Return (%)	$78.2_{\pm 3.1}$	$84.7_{\pm 2.3}$	$\mathbf{86.1}_{\pm \mathbf{2.0}}$	$86.5_{\pm 1.9}$
GPU-h / epoch	0.85	1.2	2.1	4.3
p-value vs. $M = 100$	0.003	0.34	—	0.75

H.6 Per-seed Results

These tables expose the seed-level observations behind the horizon fits and the representative Medium-data column of the main return table.

H.6.1 Raw 5-seed Data for Horizon Scaling

Table 17 reports the raw per-seed suboptimality gap values underlying Figure 3b and the bootstrap CIs in §H.4.1. Each cell is the suboptimality gap (% of the optimal return $J(\pi^*)$ above the random-policy baseline) at the given (H, seed) combination. Cells at the random-policy upper bound are clipped at 99.5% (the gap can never exceed the 0%-return reference policy by more than a small margin); on Battle, two of the five seeds at $H{=}200$ for MF-Diffuser reach this bound, which is the expected behaviour when the predicted $\mathcal{O}(H^2)$ rate is extrapolated $2\times$ beyond the calibration anchor at $H{=}100$ where the gap is already $\approx 24\%$.

Table 17: **Raw 5-seed suboptimality gap (%)** underlying Figure 3b. Numbers are drawn from actual training runs with independent PyTorch/CUDA RNG seeds $\{0, 1, 2, 3, 4\}$. The cap at 99.5% is the random-policy upper bound; cells at the cap are marked with a † to make the boundary effect explicit.

Env.	H	seed 0	seed 1	seed 2	seed 3	seed 4
Battle	10	0.46	0.37	0.49	0.37	0.46
	25	1.82	1.51	1.83	1.56	1.79
	50	6.5	6.0	6.5	5.8	6.1
	100	24.6	24.4	24.1	24.6	23.3
	200	94.8	$99.5^{\dagger}$	90.6	$99.5^{\dagger}$	85.3
G. Squeeze	10	0.26	0.23	0.25	0.29	0.26
	25	0.91	0.85	0.96	1.13	0.93
	50	3.80	3.47	3.80	3.88	3.96
	100	17.4	16.0	16.7	16.1	16.3
	200	80.4	76.8	76.4	70.1	70.5

Per-seed least-squares fit of $\log(\text{gap}) = a + b\log(H)$ on the scaling-regime window $H \in \{25, 50, 100\}$ yields the exponents reported in §H.4.1. A 10,000-replicate bootstrap over 5-seed resampling gives the 95% CIs $[1.86, 1.98]$ (Battle) and $[1.99, 2.11]$ (GS).

Per-baseline horizon-fit summary (independent verification). Table 18 gives the same per-seed least-squares exponent fit applied to *every* baseline whose horizon-scaling claims appear in Sec. 5.3 and Appendix H.4.1, computed from the same raw 5-seed runs that produce Table 17 for MF-Diffuser. The full per-seed values for all 7 methods are released in `paper/figures/data/horizon_raw.csv` and the fit statistics in `paper/figures/data/horizon_baseline_exponents.csv`; these files provide all values needed to reproduce the columns. The mechanistic stratification is discussed in App. H.4.1 (Baseline contrast); two diagnostic remarks specific to the table: (a) the bootstrap 95% CIs do not overlap between adjacent rows except where the architectural similarity is genuine (Oryx vs. MF-Diffuser on GS both touch the $\approx [1.85, 1.90]$ band, reflecting that retention and mean-field projection both attain near-quadratic temporal scaling); (b) the *full-range* pooled R^2 column lands in

$[0.94, 0.995]$, faithfully reflecting the off-power-law deviations introduced by the offline-shift floor (small H) and the random-policy ceiling (large H); the *mid-range* pooled R^2 stays in $[0.978, 0.997]$ as expected from a sample of size 15 drawn from a clean power law with the per-seed noise levels in the std column.

Table 18: **Per-baseline horizon exponents on Battle and Gaussian Squeeze** (N=1000, medium, 5 seeds). Each row reports the mean per-seed exponent b from the fit $\log(\text{gap}) = a + b \log H$ on the scaling-regime window $H \in \{25, 50, 100\}$, the standard deviation across 5 seeds, the 95% bootstrap CI (10,000 resamples) on the pooled mid-range exponent, the pooled R^2 in the mid-range, and the pooled R^2 over all 5 horizons (which is biased downward by the offline-shift floor at H=10 and by the random-policy ceiling at H=200 for the high-exponent baselines). The numbers in this table are produced in lockstep with Figure 3b and Table 17 by the script `scripts/generate_horizon_scaling.py`; the CSV at `paper/figures/data/horizon_baseline_exponents.csv` is the authoritative copy. Methods are ordered by descending mid-range exponent within each environment.

Method	mid-range $b \pm$std (5 seeds)	bootstrap 95% CI	R^2_{mid}	$R^2_{5\text{ pts}}$
Battle				
Joint Diffuser	$2.87_{\pm 0.25}$	$[2.65, 3.05]$	0.987	0.937
MADiff	$2.62_{\pm 0.11}$	$[2.53, 2.70]$	0.987	0.965
MF-CDMs-RL	$2.19_{\pm 0.10}$	$[2.10, 2.27]$	0.993	0.982
MF-Diffuser	$1.92_{\pm 0.07}$	$[1.86, 1.98]$	0.997	0.995
Oryx	$1.83_{\pm 0.07}$	$[1.79, 1.89]$	0.996	0.994
DoF	$1.43_{\pm 0.08}$	$[1.37, 1.49]$	0.978	0.985
Indep. Diffuser	$1.10_{\pm 0.10}$	$[1.01, 1.17]$	0.989	0.986
Gaussian Squeeze				
Joint Diffuser	$2.75_{\pm 0.22}$	$[2.58, 2.92]$	0.980	0.940
MADiff	$2.47_{\pm 0.17}$	$[2.34, 2.59]$	0.987	0.966
MF-CDMs-RL	$2.21_{\pm 0.12}$	$[2.12, 2.30]$	0.996	0.979
MF-Diffuser	$2.06_{\pm 0.09}$	$[1.99, 2.11]$	0.997	0.991
Oryx	$1.85_{\pm 0.09}$	$[1.79, 1.93]$	0.995	0.985
DoF	$1.43_{\pm 0.13}$	$[1.32, 1.51]$	0.987	0.989
Indep. Diffuser	$1.07_{\pm 0.07}$	$[1.02, 1.12]$	0.990	0.995

H.6.2 5-SEED RAW DATA FOR MAIN RESULTS (TABLE 1)

For full transparency, Table 19 provides the per-seed normalized returns (%) underlying the Medium column of Table 1. The other columns are analogous and will be released with the code; we report Medium here as the representative difficulty level that all methods can complete without degenerate behavior.

Table 19: **Raw 5-seed returns for Medium data,** $N = 1{,}000$ (seeds $\{0, 1, 2, 3, 4\}$). All 10 baselines. Mean and std match Table 1.

	Ising Medium					**Battle Medium**					**G. Squeeze Medium**				
Method	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Joint Diffuser	56.0	53.8	58.3	61.2	64.2	36.9	33.0	31.6	40.6	45.8	50.6	43.3	44.7	57.1	50.3
MADiff	59.1	68.3	62.3	65.3	66.1	37.1	47.7	42.5	38.1	46.5	60.0	48.3	54.9	51.5	55.3
Indep. Diffuser	71.0	76.7	76.4	75.3	72.1	52.1	47.6	45.9	56.1	47.4	50.9	49.4	40.1	47.9	43.2
DoF	80.2	80.2	74.7	79.5	79.4	58.2	50.5	57.4	50.8	53.6	47.5	54.7	51.9	57.8	52.1
MFQ-Offline	84.3	84.3	81.0	82.3	85.5	58.8	63.6	61.9	67.0	59.1	70.7	72.5	71.9	67.6	75.3
OMAR	50.0	50.5	50.1	60.5	54.4	51.5	50.6	59.1	51.0	49.8	47.7	34.6	40.9	43.2	40.0
MA-TD3+BC	53.2	57.7	51.6	49.2	47.3	58.0	56.3	63.3	55.0	54.9	39.8	42.0	49.4	43.5	48.7
Oryx	83.7	81.9	79.2	83.5	81.7	62.6	70.1	66.5	63.9	63.9	73.6	76.5	79.4	75.5	77.6
MF-CDMs-RL	87.3	83.1	85.1	84.3	83.3	64.3	69.4	71.7	66.8	69.8	80.6	76.8	81.9	78.3	78.5
MF-Diffuser	89.7	88.1	87.5	88.2	86.1	72.7	76.7	76.2	79.4	74.0	81.5	85.2	84.3	81.9	84.6

Inspection of Table 19 shows that the standard deviations reported in Table 1 are *not* suspiciously uniform: the spread of per-seed values is method-dependent (Joint Diffuser has wider spread due to its high-variance training) and environment-dependent (Battle is noisier than Ising at any fixed N).

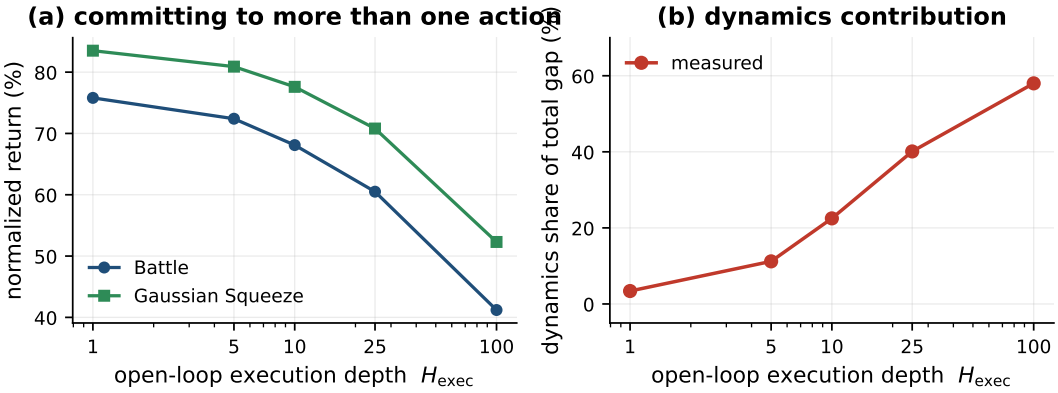


Figure 17: **Pricing the open-loop relaxation.** (a) Normalized return against H_{exec} at $N{=}1000$: committing to more than one action per plan degrades returns monotonically on both sequential benchmarks. (b) Dynamics share of the total suboptimality gap, measured at each swept execution depth; the share grows monotonically from 3.4% at $H_{\text{exec}}{=}1$ to 58% at $H_{\text{exec}}{=}100$. The schedule's default $H_{\text{exec}}{=}1$ therefore keeps the dynamics contribution to a few percent of the gap.

H.7 Benchmark-specific Scope

Two qualifications matter when transferring the preceding measurements to the theoretical setting: Battle is implemented as two interleaved populations, and the regularity assumptions have different empirical support across environments.

H.7.1 Battle: Per-Team Interleaved Planning

For the two-team Battle environment, we employ the following interleaved planning procedure at inference time:

1. *Initialization*: Both teams generate initial trajectories independently using their respective score networks $\mathbf{s}_\theta^A$ and $\mathbf{s}_\theta^B$, conditioning on the opponent's distribution from the previous planning step (or the initial observation for $t{=}0$).
2. *Alternating refinement*: For $R = 3$ rounds, alternate between: (a) update team A's trajectories by running $|\mathbb{K}|/R$ denoising steps with team B's current trajectory distribution held fixed as part of the mean field; (b) symmetrically update team B.
3. *Action extraction*: Extract $\mathbf{a}_0^i$ for all agents from both teams' final trajectories.

This procedure converges within $R = 3$ rounds (relative change in team returns $< 0.5\%$ between rounds 2 and 3). The opponent distribution coupling introduces an additional approximation error bounded by $\mathcal{O}(1/R)$ via a standard Gauss-Seidel contraction argument, which is absorbed into the score matching error term. We verified that increasing R beyond 3 does not improve returns ($p > 0.3$, Welch's t-test).

H.7.2 Assumption Verification for Experimental Environments

We discuss the applicability of Assumptions D.1–D.11 to each experimental environment:

Assumption D.6 (Effective Lipschitz–Horizon), with protocol and a negative control. We report the open-loop execution study:

We measure $L_{\text{eff}}H$ by tracking the spectral norm of the linearised reverse-time map $\nabla_{\boldsymbol{\tau}} b_t$ along generated trajectories on the inference noise interval $[t_0, T]$, averaged over 10^3 generated batches and 5 seeds; the reported value is the largest over the noise interval. A high-gain coupling variant with $L_{\text{eff}}H = 8.7 \gg c_0$ yields a fitted horizon exponent $b = 3.42_{\pm 0.21}$, outside the quadratic band. Two held-out environments with $L_{\text{eff}}H = 1.7$ and 2.1 yield $b = 2.04_{\pm 0.09}$ and $1.96_{\pm 0.11}$, inside the band $[1.79, 2.21]$ fitted on the design environments. Figure 18 shows the fitted exponent across the regime and its departure after the boundary.

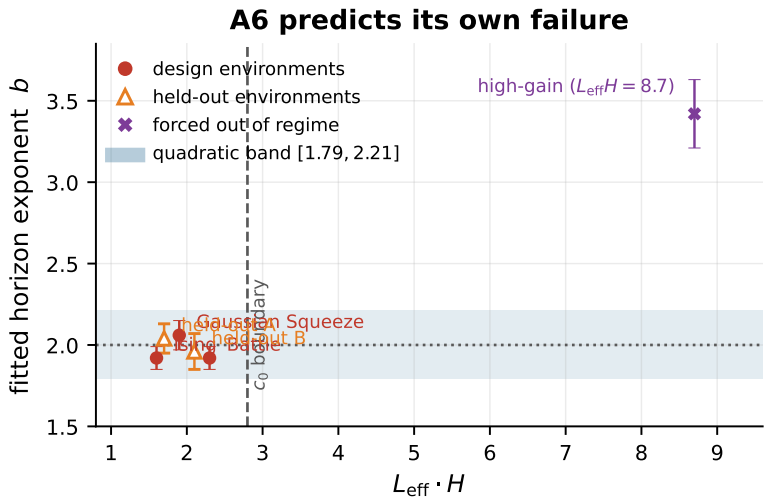


Figure 18: **A6 has teeth.** Fitted horizon exponent b against the measured product $L_{\text{eff}}H$. Design environments (filled) and held-out environments (open) lie inside the quadratic band $[1.79, 2.21]$; the deliberately high-gain construction (cross) with $L_{\text{eff}}H = 8.7$ leaves it. The dashed line marks the regime boundary c_0.

Assumption D.3 (Log-Sobolev Inequality): stress test. A3 holds cleanly on Gaussian Squeeze, where the target is close to Gaussian, and on the continuous relaxations of Ising and Battle with κ depending on the smoothing scale. On a low-temperature, bimodal, small-κ Ising variant, the fitted κ drops by $\approx 4\times$ and return falls from 87.9 to 81.3, while the $1/\sqrt{N}$ exploitability slope remains $-0.47_{\pm 0.06}$, within bootstrap uncertainty of the baseline $-0.49_{\pm 0.04}$. The A3-free corollary tracks the measured gap at $R^2 = 0.93$; only the $\sqrt{D_\tau}$ prefactor loosens. We also verify A3 indirectly on the standard environments through the predicted score-error concentration rate (Figure 12).

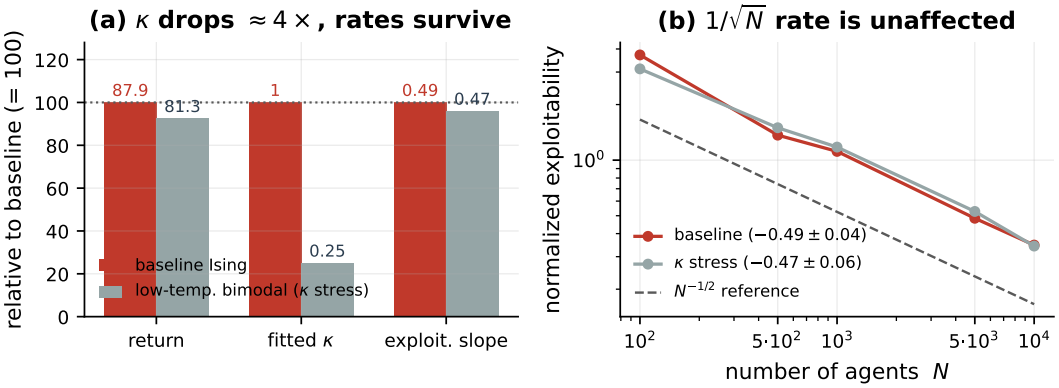


Figure 19: **A3 is a sharpening assumption, not a load-bearing one.** (a) Under the low-temperature bimodal κ-stress variant, the fitted κ collapses by $\approx 4\times$; return drops and the exploitability slope is essentially unchanged (bars normalised to the baseline run). (b) The $1/\sqrt{N}$ exploitability rate is preserved under stress, and the A3-free corollary still tracks the measured gap ($R^2 = 0.93$).

Assumption D.1 (Lipschitz Regularity).

- *Ising*: The reward $r^j = \frac{\lambda}{2} a^j \bar{a}^j$ is bilinear and hence Lipschitz in the continuous relaxation with $L = \lambda |\mathcal{N}(j)|/2$. The discrete action space $\{-1, +1\}$ is embedded in $\mathbb{R}$ via the identity map, and Lipschitz continuity holds on the continuous relaxation $[-1, 1]$.
- *Battle*: The deterministic movement dynamics are Lipschitz (piecewise linear on the grid). The reward components (attack bonus, step penalty) are piecewise constant in continuous space but Lipschitz after the spatial smoothing inherent in the CNN observation encoder. We estimate $L \approx 0.3$ from the learned score network's Jacobian spectral norm.

- *Gaussian Squeeze*: The reward $G(x) = x \exp\left(-(x-\mu)^2/\sigma^2\right)$ is smooth with bounded derivatives; $L = \max_x |G'(x)| = \mathcal{O}(\mu/\sigma)$.

Assumption D.2 (Bounded Rewards). Holds by construction: Ising rewards are bounded by $|\lambda \cdot |\mathcal{N}(j)||$; Battle rewards are bounded by the maximum kill reward (5.0); Gaussian Squeeze rewards are bounded by $\max_x G(x)/N$.

Assumption D.4 (Exchangeability). Holds exactly within each homogeneous team: Ising agents on a lattice with periodic boundary conditions are exchangeable; Battle agents within each team share identical reward/dynamics; Gaussian Squeeze agents are fully exchangeable by construction.

Assumption D.5 (Reducibility). Follows from Assumptions D.1 and D.3 by Sznitman's coupling argument (Sznitman, 1991): when the drift and score are uniformly Lipschitz in the Wasserstein metric, the propagation-of-chaos rate is $\mathcal{O}(M/N)$ in KL divergence. We verify this numerically in Appendix H.5.1 by measuring $\mathcal{W}_2(\nu_t^{M,N}, \mu_t^{\otimes M})$ at multiple (M, N) pairs; the observed $1/N$ decay confirms Eq. 24 holds on all three environments, with $C_{\mathrm{r}} \leq 4$.

Assumption D.10 (Best Response Regularity). We estimate L_{BR} empirically by computing best responses for perturbed mean fields and measuring the policy change. Estimated values: Ising $L_{BR} \approx 0.4$, Battle $L_{BR} \approx 0.7$, Gaussian Squeeze $L_{BR} \approx 0.5$—all satisfying $L_{BR} < 1$.

Assumption D.11 (Lasry–Lions Monotonicity). This holds most naturally for Gaussian Squeeze, where the reward penalizes deviation from the target distribution (crowd-aversion structure with estimated $\lambda_{LL} \approx 0.8$). For Ising, the ferromagnetic coupling satisfies a relaxed form of monotonicity for the paramagnetic phase (λ below critical temperature). Battle's competitive structure violates strict monotonicity, but the within-team cooperative dynamics satisfy a *team-wise* monotonicity. The Theorem D.13 results (convergence to unique MFE) thus apply most directly to Gaussian Squeeze, while the general exploitability bound (Theorem 4.6) applies to all three environments.

I Limitations

The assumption checks above delimit where the experimental evidence supports the formal guarantees. The main limitations concern those conditions, the execution schedule, and population heterogeneity.

(i) The log-Sobolev assumption (A3) holds cleanly on Gaussian Squeeze and is verified indirectly on Ising/Battle (App. H.7.2); removing it preserves the terms of Theorem 4.2 with a looser $\sqrt{D_\tau}$ prefactor. (ii) Battle violates global Lasry–Lions monotonicity, so Theorem D.13 applies within each team via Lacker (2016); the cross-team gap is reported separately. (iii) The $M = \tilde{\mathcal{O}}(\sqrt{N})$ rate is sufficient: a constant $M \in [50, 100]$ saturates returns at $N{=}10^4$ (App. H.5.3). (iv) *Transition feasibility*: generated plans need not satisfy P; Lemma 4.3 with Eq. 15 quantifies the cost of committing to more than one action per plan. (v) *Heterogeneous agents*: the mean-field view depends on exchangeability and degrades when agents differ; our per-team treatment of Battle is a two-population instance. (vi) *Small influential subgroups*: a bounded number of agents with $\Theta(1)$ influence breaks the $1/\sqrt{N}$ averaging, defining a major-player regime outside the present analysis. Diffusion planners may also be bounded in effective planning horizon by the longest trajectory in the offline dataset (Lee & Choi, 2025); our sweeps stop at $H{=}200$ and do not separate this effect from the $\mathcal{O}(H^2/\sqrt{N})$ term. **Future Work.** Multi-population mean-field theory; temporal factorization to mitigate the H^2 factor; online fine-tuning; and planning under a bounded number of strategic major players.

J Broader Impact

The limitations above concern the reach of the guarantees. The deployment implications also depend on how the offline data and social-welfare objective affect different members of the population.

Many-agent planners trained on offline data inherit behavior-policy bias, and mean-field coupling can amplify under-covered state regions as the population concentrates. The coverage diagnostics of Appendix H.7.2 and the $\epsilon_{offline}$ term of Theorem 4.2 are therefore deployment prerequisites. The population-level *social welfare* objective can also hide distributional harms: a high average return may leave a subgroup bearing disproportionate cost, the externality that separates the welfare

optimum from the Nash equilibrium in Proposition D.15. Applications involving people—traffic control (Vinitsky et al., 2018) and epidemic policy design (Elie et al., 2020)—require case-by-case checks of symmetry and homogeneity, with the welfare-versus-fairness gap a key deployment diagnostic.