DIFFUSION WORLD MODELS FOR TEMPORAL OBJECTIVE ALIASING IN DECENTRALIZED MARL

Anonymous authors
Paper under double-blind review

ABSTRACT

World models for multi-agent reinforcement learning (MARL) face a fundamental ambiguity under partial observability: the same local observation can correspond to multiple long-horizon objectives, inducing distinct future trajectory distributions. We refer to this phenomenon as *temporal objective aliasing*. Unlike conventional perceptual aliasing, which arises from ambiguity in the underlying state, temporal objective aliasing arises when historical interaction contexts imply different objectives despite similar current observations. Consequently, a world model conditioned only on local observations may collapse multiple future modes, leading to inaccurate imagined trajectories and suboptimal policy updates. Moreover, centralized multi-agent world models face combinatorial growth in joint state-action spaces, whereas purely local models may overlook coordination-relevant interactions and yield suboptimal decisions. To address these challenges, we propose **MODAL** (**M**odel-based **O**bjective-aware **D**iffusion for M**ARL**), a decentralized framework where each agent models local long-horizon trajectories with a diffusion world model and summarizes neighborhood interactions through weighted mean field (MF) communication. A history encoder further infers latent objectives from local histories to guide trajectory prediction and policy optimization. We establish theoretical guarantees on MODAL's ability to preserve continuation-value distinctions under temporal objective aliasing. Experiments on MOSMAC and Flatland show that MODAL outperforms established MARL baselines, with larger gains as team size and coordination complexity grow.

1 INTRODUCTION

Most real-world cooperative multi-agent tasks require agents to coordinate over long horizons while adapting to multiple evolving objectives (Li et al., 2026; Nayak et al., 2024; Sivagnanam et al., 2024; Rother et al., 2025), yet agents can only observe scalar utilities aggregated using fixed scalarization weights. Figure 1 illustrates this challenge in MOSMAC (Geng et al., 2025), where agents cannot access individual reward components as trajectories progress through different task stages. Under partial observability, however, similar local observations can arise at different progress stages: an unseen enemy may be either out of view or already eliminated, while the same relative distance may correspond to approaching or retreating dynamics. We refer to this history-dependent ambiguity between local observations and implicit behavioral objectives as *temporal objective aliasing*.

Definition 1 (Temporal Objective Aliasing)**.** Under fixed reward scalarization, temporal objective aliasing is a task-structured form of perceptual aliasing: similar observations from distinct histories imply different continuation-value profiles and potentially different action preferences.

To address this challenge, most multi-agent reinforcement learning (MARL) methods learn cooperative policies through repeated trial-and-error environment interactions (Hu et al., 2023; Li et al., 2026; Na & Moon, 2024), which can be data-intensive and require unsafe exploration in complex systems. World model-based reinforcement learning (Hafner et al., 2025; Alonso et al., 2024; Chen et al., 2022; Ghugare et al., 2023) offers a natural alternative by learning a predictive environment dynamics model and using it to simulate future interactions, thereby reducing the reliance on online interaction. Yet extending this paradigm to multi-agent settings is nontrivial (Zhao et al., 2026; Zhang et al., 2025c;a; Duan et al., 2026). Under centralized training with decentralized execution (CTDE) (Zhang et al., 2025b;c; Liu et al., 2024; Zhu et al., 2024), centralized world models exploit

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

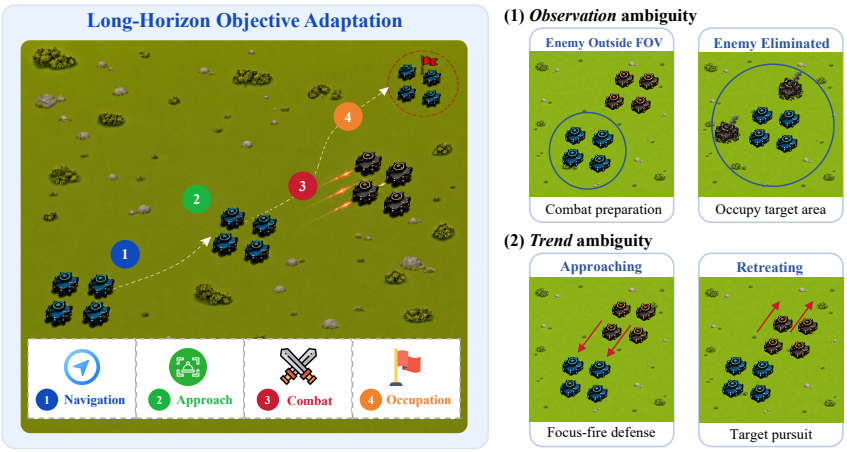


Figure 1: **Temporal objective aliasing in MOSMAC.** Left: a typical episode progresses through four behavioral stages: navigation, approach, combat, and occupation. Right: two representative examples of temporal objective aliasing: observation ambiguity and trend ambiguity.

global states or joint actions but scale poorly with team size, whereas decentralized training with decentralized execution (DTDE) (Zhao et al., 2026; Zhang et al., 2021; Takayama et al., 2026) improves scalability through local modeling but may overlook neighborhood interactions and degrade both dynamics prediction and coordinated decisions. This motivates a scalable decentralized world model that preserves coordination-relevant information without high-dimensional global inputs.

Beyond scalable decentralized modeling, mitigating temporal objective aliasing requires capturing history-conditioned, multimodal long-horizon futures. Existing objective-aware world models (Zheng et al., 2026; Mazzaglia et al., 2022; Zhao et al., 2025; Wang et al., 2026), however, largely target single-agent settings and rely on large offline datasets covering predefined objective conditions. This dependence imposes strong demands on dataset scale and coverage, while limiting adaptation to objectives that evolve during interaction. Moreover, these methods do not explicitly address the coupled dynamics and coordination dependencies of decentralized multi-agent systems.

To address these challenges, we propose **MODAL**, a **M**odel-based **O**bjective-aware **D**iffusion for **MARL** under DTDE. Rather than modeling centralized joint dynamics, MODAL models long-horizon multimodal futures with local diffusion world models, while weighted mean field (MF) communication (Hao et al., 2023) captures neighborhood interactions, and a history encoder infers latent objectives from local interaction histories to align imagined trajectories with the decision stage. We theoretically analyze how diffusion world modeling in MODAL mitigates temporal objective aliasing under partial observability. Experiments on MOSMAC and Flatland demonstrate consistent gains over strong MARL baselines. Our main contributions are summarized as follows:

- We formalize *temporal objective aliasing* in partially observable MARL, where similar local observations can conceal distinct task-progress histories requiring different behaviors.
- We introduce MODAL, a DTDE framework that captures multimodal long-horizon futures with local diffusion world models and preserves coordination cues through distance-weighted MF communication. History-based objective inference further aligns imagined rollouts with evolving decision demands, enabling objective-consistent policy learning.
- Under temporal objective aliasing, we theoretically show that MODAL preserves trajectory-distribution and continuation-value separation with bounded modeling error. Experiments on MOSMAC and Flatland demonstrate improved performance and scalability.

2 Related Work

History-Based Context Representation. In long-horizon multi-agent coordination, similar observations may correspond to distinct task stages, motivating the use of historical context. Accordingly, recurrent and Transformer-based encoders aggregate local histories (Omidshafiei et al., 2017;

Hu et al., 2021), while masked contrastive learning captures temporal and inter-agent dependencies (Song et al., 2023). Other approaches represent agents' behavior (Papoudakis et al., 2021a) or emergent roles (Hu et al., 2024b). At the task level, inferred contexts and trajectory classes support continual coordination and multi-task generalization (Yuan et al., 2025; Na et al., 2025). However, these context-aware approaches predominantly focus on supporting model-free decision-making, leaving history-conditioned trajectory generation for imagined policy optimization underexplored.

Multi-Agent World Models. World models learn predictive dynamics to generate imagined trajectories for data-efficient policy learning, commonly using recurrent state-space model (RSSM)-based (Hafner et al., 2019; 2020; 2025), Transformer-based (Chen et al., 2022; Robine et al., 2023; Zhang et al., 2023), or diffusion-based (Alonso et al., 2024; Rigter et al., 2024; Lee et al., 2025; Li et al., 2025b) architectures. In multi-agent settings, MAMBA (Egorov & Shpilman, 2022) augments RSSMs with latent communication, whereas DMAWM (Xue et al., 2026) factorizes agent-specific and shared dynamics. For Transformer-based modeling, MARIE (Zhang et al., 2025b) uses an autoregressive Transformer with centralized aggregation. Diffusion-based MADiff (Zhu et al., 2024) models multi-agent trajectories with attention, whereas DIMA (Zhang et al., 2025c) formulates state transitions through agent-wise diffusion. However, these methods predominantly follow CTDE and rely on global states, joint actions, or centralized representations, limiting scalability as the number of agents increases.

Inter-Agent Communication. Although DTDE offers favorable scalability, isolated local learning can limit coordination and exacerbate perceived non-stationarity (Omidshafiei et al., 2017; Gerstgrasser & Parkes, 2022). Communication-based MARL learns task-relevant message exchange (Lo et al., 2024; Wang et al., 2025), while graph-based methods optimize communication topology and information propagation (Hu et al., 2024a; Li et al., 2025a). However, their agent- or edge-level processing overhead grows with the communication graph. Mean field (MF) communication (Yang et al., 2018; Cui et al., 2024) instead compresses neighborhood interactions into fixed-dimensional statistics by coupling each agent with the aggregate effect of its neighbors. However, standard mean aggregation treats neighbors uniformly and may obscure heterogeneous influence. Instead, MODAL combines distance-weighted aggregation with neighbor-presence and density statistics. We further validate its effectiveness empirically and theoretically for mitigating temporal objective aliasing.

3 PRELIMINARIES

3.1 MULTI-AGENT SYSTEMS AS DEC-POMDPS

We model a fully cooperative multi-agent system as a decentralized partially observable Markov decision process (Dec-POMDP) (Geng et al., 2025), defined by $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, P, \{\Omega^i\}_{i=1}^n, \boldsymbol{\mathcal{O}}, \{\mathbf{R}^i\}_{i=1}^n, \boldsymbol{\omega}, \gamma \rangle$. Here, $\mathcal{N} = \{1, \dots, n\}$ is the agent set, $\mathcal{S}$ is the global state space, and $\boldsymbol{\mathcal{A}} = \prod_{i=1}^n \mathcal{A}^i$ is the joint action space. The transition function $P(s_{t+1} \mid s_t, \mathbf{a}_t)$ governs the dynamics under joint action $\mathbf{a}_t = (a_t^1, \dots, a_t^n)$, while each observation function $\mathcal{O}^i : \mathcal{S} \to \Omega^i$ maps the global state to agent i's local observation. The environment defines a vector reward $\mathbf{r}_t^i = \mathbf{R}^i(s_t, \mathbf{a}_t) \in \mathbb{R}^d$, but agent i only observes the scalarized utility $u_t^i = \boldsymbol{\omega}^\top \mathbf{r}_t^i$, where $\boldsymbol{\omega}$ is a fixed scalarization weight vector. Under decentralized execution, each policy π^i uses only locally available information, and the joint policy $\boldsymbol{\pi}$ maximizes $J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}[\sum_{t=0}^{\infty} \gamma^t \frac{1}{n} \sum_{i=1}^n u_t^i]$.

A fixed scalarization specifies the overall utility preference, yet the appropriate action preferences can vary with task progress. Let $\mathcal{E}_t^i$ denote agent i's local interaction history. For each current action $a \in \mathcal{A}^i$, define its H-step continuation value as

$$Q_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i, a) = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{h=0}^{H-1} \gamma^h u_{t+h}^i \,\middle|\, \mathcal{E}_t^i, a_t^i = a \right]. \tag{1}$$

Let $\mathbf{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i) = [Q_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i, a)]_{a \in \mathcal{A}^i}$ be the continuation-value profile. With a fixed reward preference, distinct histories may encode task-progress regimes having different action preferences. Given thresholds $\epsilon_o, \epsilon_Q > 0$, temporal objective aliasing in Definition 1 occurs when

$$\left\| o_t^i - o_{t'}^i \right\|_2 \le \epsilon_o, \qquad \left\| \mathbf{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i) - \mathbf{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_{t'}^i) \right\|_\infty \ge \epsilon_Q. \tag{2}$$

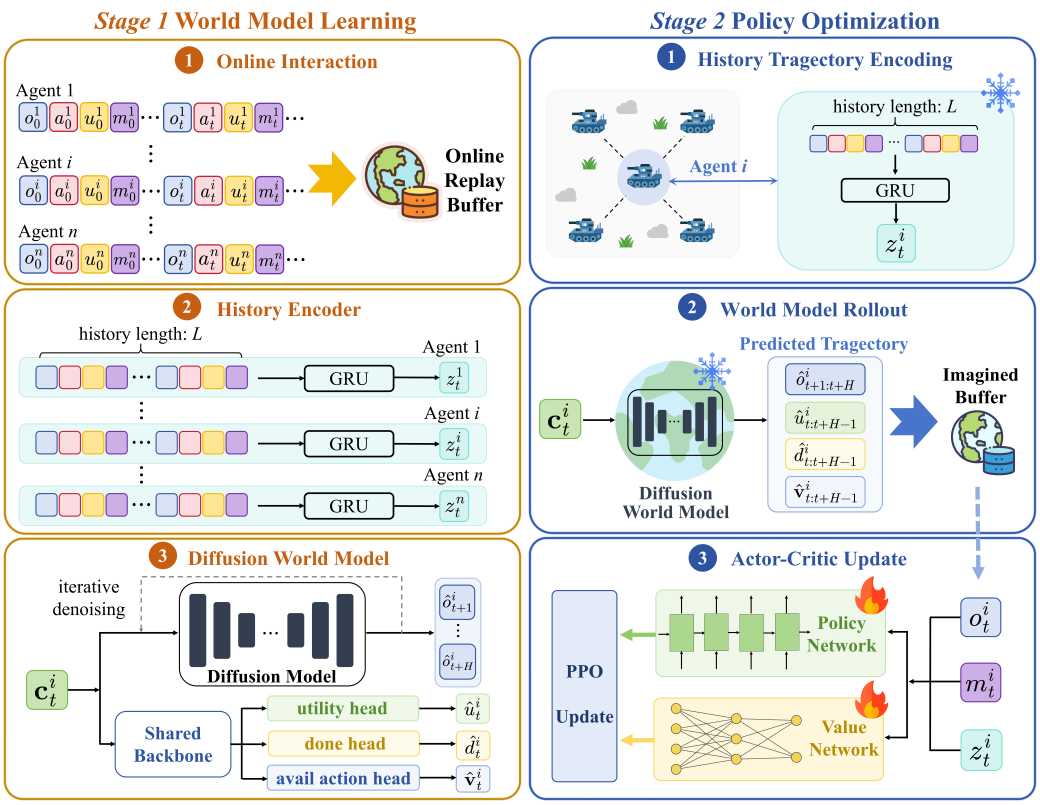


Figure 2: **Overview of MODAL.** Left: Stage 1 learns history representations and a local diffusion world model from online trajectories augmented with weighted MF messages. Right: Stage 2 generates imagined future trajectories for decentralized actor-critic policy optimization.

3.2 Score-Based Diffusion Models

Diffusion models generate samples by reversing a progressive noising process (Sohl-Dickstein et al., 2015; Song et al., 2021). We use $\tau \in [0, T]$ to denote diffusion time and reserve t for environment time. Given a clean sample $\mathbf{x}(0) \sim p_{\text{data}}$ from the dataset, a noisy sample is obtained as

$$\mathbf{x}(\tau) = \alpha(\tau)\mathbf{x}(0) + \sigma(\tau)\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $\alpha(\tau)$ and $\sigma(\tau)$ control the signal and noise scales, respectively. Using the standard noise prediction objective (Ho et al., 2020), a conditional neural noise model $\boldsymbol{\epsilon}_\theta(\mathbf{x}(\tau), \tau, \mathbf{c})$ is trained as

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{\mathbf{x}(0),\tau,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}(\tau), \tau, \mathbf{c})\|_2^2\right], \tag{4}$$

where $\mathbf{c}$ denotes the conditioning information. At inference, the learned reverse process starts from Gaussian noise and iteratively generates samples from the conditional data distribution $p_\theta(\mathbf{x}(0) \mid \mathbf{c})$.

4 Method

To mitigate temporal objective aliasing while preserving decentralized scalability, we propose **MODAL**, illustrated in Figure 2. Under DTDE, MODAL alternates between *world model learning* and *policy optimization*. Each agent first summarizes neighborhood interactions through weighted MF communication and extracts an objective-relevant representation from interaction history. Conditioned on these representations, a local diffusion world model captures multimodal long-horizon dynamics and generates imagined experience. The resulting imagined rollouts are incorporated into an actor-critic framework for policy learning, improving sample efficiency during online interaction.

4.1 Weighted MF Communication

Under DTDE, purely local world models fail to capture inter-agent dependencies, whereas explicit conditioning on all neighbors yields variable-size inputs. Standard MF aggregation (Yang et al., 2018) provides a fixed-dimensional summary but obscures heterogeneous neighbor influence through uniform averaging. As illustrated in Figure 3, MODAL instead combines distance-weighted aggregation with neighbor-presence and density statistics.

Specifically, agent i communicates only with alive agents within its view radius R_{view}. Its observable neighborhood at timestep t is

$$\mathcal{C}_t^i = \{j \in \mathcal{N} \setminus \{i\} \mid d_{ij} \le R_{\text{view}},\ j \text{ is alive}\}, \quad (5)$$

where d_{ij} is the pairwise distance. For nonempty neighborhoods, we prioritize nearby agents using

$$w_{ij} = \frac{\exp(-\kappa d_{ij})}{\sum_{k \in \mathcal{C}_t^i} \exp(-\kappa d_{ik})}, \qquad j \in \mathcal{C}_t^i, \quad (6)$$

where $\kappa > 0$ controls the distance decay. Therefore, the weighted neighborhood aggregate is

$$\bar{o}_t^i = \sum\nolimits_{j \in \mathcal{C}_t^i} w_{ij} o_t^j. \quad (7)$$

For empty neighborhoods, we set $\bar{o}_t^i = \mathbf{0}$. Since normalization discards cardinality, we further include presence and normalized density indicators:

$$h_t^i = \mathbb{I}\left(|\mathcal{C}_t^i| > 0\right), \qquad \rho_t^i = \frac{|\mathcal{C}_t^i|}{n}. \quad (8)$$

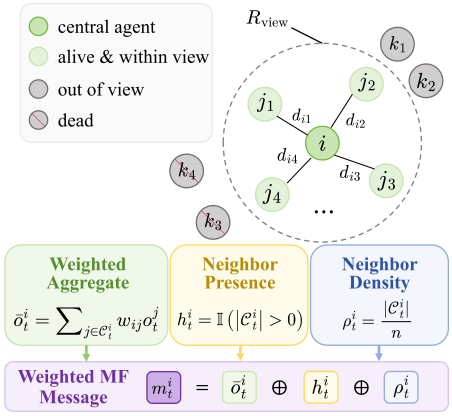


Figure 3: **Weighted MF communication.** Each agent aggregates observations from visible alive neighbors using distance-based weights prioritizing nearby agents. The aggregate is combined with presence and density statistics to form a fixed-dimensional MF message.

The complete weighted MF message is

$$m_t^i = \bar{o}_t^i \oplus h_t^i \oplus \rho_t^i, \quad (9)$$

where $\oplus$ denotes concatenation. Thus, m_t^i remains fixed-dimensional across team sizes while preserving heterogeneous neighborhood effects for multi-agent collaboration.

4.2 History-Based Objective Representation

To mitigate temporal objective aliasing, each agent infers an objective-relevant representation from recent interaction history rather than relying solely on the current observation. Specifically, agent i encodes its length-L observation-action-utility sequence together with the weighted MF messages:

$$z_t^i = f_\psi\left(\{[o_\ell^i, a_\ell^i, u_\ell^i, m_\ell^i]\}_{\ell=t-L}^{t-1}\right), \quad (10)$$

where f_ψ is a recurrent history encoder and z_t^i denotes the latent objective representation. The encoder is trained end-to-end with the diffusion world model introduced in Section 4.3. In particular, z_t^i provides history-informed context for subsequent trajectory prediction and actor-critic learning, thereby aligning imagined dynamics and policy updates with the underlying objective.

4.3 Objective-Conditioned Diffusion World Model

Conditioning on z_t^i resolves history-dependent objective context, yet stochastic interactions may still induce multiple plausible long-horizon futures. Deterministic models tend to average plausible modes, while recursive one-step prediction accumulates errors. MODAL therefore uses a conditional diffusion model to jointly generate the H-step local observation trajectory

$$\mathbf{x}_t^i(0) = [o_{t+1}^i, \cdots, o_{t+H}^i], \qquad \mathbf{c}_t^i = [o_t^i, a_t^i, m_t^i, z_t^i]. \quad (11)$$

Following the noise prediction formulation in Eq. (4), the trajectory model is trained by

$$\mathcal{L}_{\text{traj}} = \mathbb{E}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t^i(\tau), \tau, \mathbf{c}_t^i\right)\right\|_2^2\right], \quad (12)$$

where $\mathbf{x}_t^i(\tau)$ is obtained from Eq. (3). The multimodal capability of diffusion models enables trajectory generation to capture diverse long-horizon outcomes, supporting coherent planning.

Observation trajectories alone are insufficient for imagined policy learning. MODAL therefore introduces auxiliary heads that predict utility u, the termination indicator d, and the available-action mask $\mathbf{v}$, which respectively provide optimization signals, prevent invalid continuation, and constrain feasible actions. Their joint loss is

$$\mathcal{L}_{\text{aux}} = \lambda_u \mathcal{L}_{\text{util}} + \lambda_d \mathcal{L}_{\text{term}} + \lambda_a \mathcal{L}_{\text{avail}}. \tag{13}$$

The history encoder, diffusion noise model, and auxiliary heads are trained end-to-end using

$$\mathcal{L}_{\text{WM}} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{aux}}. \tag{14}$$

Together, these predictions enable MODAL to construct complete objective-conditioned imagined rollouts that capture long-horizon local dynamics.

4.4 Policy Learning

After learning the objective-conditioned world model from online trajectories, MODAL generates H-step rollouts to enrich the training data for actor-critic learning. Following the IPPO framework (de Witt et al., 2020), homogeneous agents share actor and critic parameters while making decisions based on decentralized information. For agent i, both networks take $\boldsymbol{\xi}_t^i = [o_t^i, m_t^i, z_t^i]$ as input. Specifically, the actor is optimized using the clipped PPO objective:

$$\mathcal{L}_\pi = -\mathbb{E}_{\mathcal{B}}\left[\min\left(r_t^i(\phi)\widehat{A}_t^i,\ \text{clip}\left(r_t^i(\phi), 1-\epsilon, 1+\epsilon\right)\widehat{A}_t^i\right)\right] - \lambda_{\mathcal{H}}\mathbb{E}_{\mathcal{B}}\left[\mathcal{H}\left(\pi_\phi(\cdot \mid \boldsymbol{\xi}_t^i)\right)\right], \tag{15}$$

where $r_t^i(\phi) = \pi_\phi(a_t^i \mid \boldsymbol{\xi}_t^i)/\pi_{\phi_{\text{old}}}(a_t^i \mid \boldsymbol{\xi}_t^i)$ is the probability ratio, $\widehat{A}_t^i$ is the advantage estimate, $\lambda_{\mathcal{H}}$ controls the entropy regularization strength, and $\mathcal{B}$ denotes the training batch. The critic augments the sampled H-step utility $\widehat{G}_{t,H}^i = \sum_{h=0}^{H-1}\gamma^h \hat{u}_{t+h}^i$ with a bootstrapped continuation value as

$$\mathcal{L}_V = \mathbb{E}_{\mathcal{B}}\left[\left(V_\nu(\boldsymbol{\xi}_t^i) - \text{sg}\left[\widehat{G}_{t,H}^i + \gamma^H V_{\bar{\nu}}(\boldsymbol{\xi}_{t+H}^i)\right]\right)^2\right], \tag{16}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator and $V_{\bar{\nu}}$ is the target critic.

During training, online trajectories are used to update the history encoder and world model while providing starting points for imagined rollouts. MODAL performs receding-horizon imagination, with the resulting imagined trajectories improving policy learning by providing additional long-horizon experience. The complete training procedure is provided in Appendix A. During execution, the decentralized policy produces coordinated decisions aligned with the current-stage objective using only local neighborhood information.

5 Theoretical Analysis

In this section, we theoretically characterize how MODAL mitigates temporal objective aliasing through history-conditioned objective representations and weighted MF communication.

Let $\mathbf{o}_t^i = \bigoplus_{j\in\mathcal{C}_t^i} o_t^j$ collect the local observations of all visible and alive neighbors, while $z_t^{i,*}$ denotes an objective-sufficient representation inferred from the complete interaction history. We assume that the finite-history inference error is uniformly bounded as $\|z_t^{i,*} - z_t^i\|_2 \le e_z$. If the conditional noise predictor is L_z-Lipschitz with respect to z, the objective-induced noise prediction error is bounded by $L_z e_z$. Following the weighted MF analysis of Hao et al. (2023) and extending the argument to the conditional diffusion predictor, $L_m(\tau)$-smoothness in neighbor observations yields an MF approximation error bounded by $L_m(\tau)C$, where C bounds the weighted neighborhood dispersion; the derivation is provided in Appendix B.1. Suppressing common diffusion input arguments for brevity, the resulting noise prediction error satisfies

$$\left\|\boldsymbol{\epsilon}^*\left(z_t^{i,*}, \mathbf{o}_t^i\right) - \boldsymbol{\epsilon}\left(z_t^i, m_t^i\right)\right\|_2 \le L_z e_z + L_m(\tau)C =: \delta_\epsilon(\tau). \tag{17}$$

Under VP-SDE (Song et al., 2021), let $\beta(\tau)$ be the noise schedule and $\bar{\alpha}(\tau) = \exp(-\int_0^\tau \beta(s)\,ds)$ the cumulative signal coefficient. For fixed local history $\mathcal{E}_t^i$ and action a, let $\mathbb{P}_H^{i,\boldsymbol{\pi}}$ and $\widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}$ denote the true and predicted distributions of the H-step local observation trajectory under $\boldsymbol{\pi}$. This yields the following bound.

Theorem 1 (Diffusion Local Trajectory Error Bound)**.** Suppose Eq. (17) holds uniformly over histories and actions. Under standard regularity conditions,

$$D_{\mathrm{KL}}\left(\mathbb{P}_H^{i,\boldsymbol{\pi}} \middle\| \widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}\right) \le \Delta_T + \frac{1}{2}\int_0^T \frac{\beta(\tau)}{1-\bar{\alpha}(\tau)}\delta_\epsilon(\tau)^2\, d\tau =: \Xi, \tag{18}$$

where Δ_T is the finite-time prior mismatch, which vanishes as the terminal forward distribution approaches the standard Gaussian. By Pinsker's inequality (Tsybakov, 2009, Lemma 2.5), their total-variation distance is at most $\sqrt{\Xi/2} =: \delta_P$.

Consequently, Theorem 1 yields a bounded error in the complete value profile.

Corollary 1 (Continuation-Value Approximation Bound)**.** Assume $0 \le \gamma < 1$, $|u_t^i| \le U_{\max}$, and define $U_H := U_{\max}(1-\gamma^H)/(1-\gamma)$. Then, for any local history $\mathcal{E}_t^i$,

$$\left\| \mathbf{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i) - \widetilde{\mathbf{Q}}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i) \right\|_\infty \le 2U_H\delta_P. \tag{19}$$

Applying this to temporally aliased histories yields a sufficient condition for value-profile separation.

Theorem 2 (Value-Profile Separation under Temporal Objective Aliasing)**.** Fix a joint policy $\boldsymbol{\pi}$. Under Theorem 1 and Corollary 1, any two histories $\mathcal{E}_t^i$ and $\mathcal{E}_{t'}^i$ satisfying Eq. (2) obey

$$\left\| \widetilde{\mathbf{Q}}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i) - \widetilde{\mathbf{Q}}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_{t'}^i) \right\|_\infty \ge \epsilon_Q - 4U_H\delta_P. \tag{20}$$

Theorem 2 follows from Corollary 1 via the reverse triangle inequality. Thus, by controlling the modeling error to satisfy $4U_H\delta_P < \epsilon_Q$, MODAL preserves a nonzero value profile separation between temporally aliased observations. Detailed proofs are provided in Appendix B.

6 Experiments

6.1 Experimental Setup

Environments. We evaluate MODAL on two cooperative multi-agent benchmarks with complementary long-horizon coordination challenges: MOSMAC (Geng et al., 2025) and Flatland (Mohanty et al., 2020). MOSMAC considers multiple objectives, including combat, safety, and navigation, and provides both single-task and sequential multi-task scenarios. We use its four single-task siege-tank scenarios, `3t`, `4t`, `8t`, and `12t`, where agents must jointly reach a randomized strategic position while coordinating navigation and combat against enemies guarding the target. Flatland models multi-train routing on constrained railway networks, where agents must reach their destinations while resolving contention on shared tracks and junctions. We evaluate the `5_agents`, `10_agents`, and `20_agents` configurations, which progressively increase traffic density and coordination difficulty. Detailed environment configurations are provided in Appendix C.

Baselines. We compare MODAL with four MARL baselines, including three model-free methods and one world model-based method. IPPO (de Witt et al., 2020) applies independent PPO with local information, IA2C (Papoudakis et al., 2021b) employs independent advantage actor-critic learning, and MAPPO (Yu et al., 2022) follows CTDE with decentralized actors and a centralized critic. For world model-based MARL, MAMBA (Egorov & Shpilman, 2022) extends DreamerV2 (Hafner et al., 2021) to multi-agent settings. To evaluate world model prediction accuracy, we additionally compare with DIMA (Zhang et al., 2025c), which employs a diffusion-inspired centralized world model with sequential conditional denoising over agents' actions. Each method is evaluated with 8 seeds per scenario, using 20 episodes per evaluation in MOSMAC and 80 in Flatland. For controlled comparison among world model-based methods, we restrict the same historical context and rollout lengths: 8 history steps and 4 rollout steps in MOSMAC, and 16 history steps and 8 rollout steps in Flatland. Additional implementation details and hyperparameters are in Appendix D.

6.2 Main Results

MODAL consistently improves sample efficiency, final performance, and scalability. As shown in Figure 4, MODAL remains competitive on simpler settings while achieving larger gains in challenging scenarios. In MOSMAC, an episode is counted as a win if at least one ally survives, all surviving allies are within the target region, and no enemy remains; Flatland success rate is the fraction

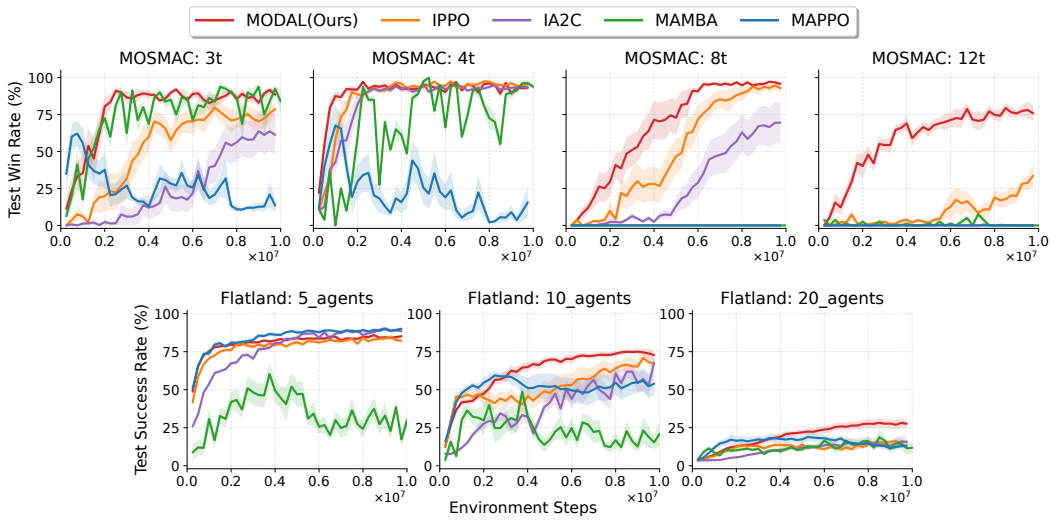


Figure 4: **Test episode win/success rates of all methods on MOSMAC and Flatland.** Curves show the mean performance over 8 random seeds, with shaded regions indicating the standard error.

of trains reaching their destinations by episode end. MODAL converges faster in `8t` and maintains a clear margin in `12t`; on Flatland, it achieves the highest success rates in `10_agents` and continues improving in `20_agents`, where competing methods degrade as coordination dependencies increase. These results indicate that model-free baselines suffer from limited interaction efficiency, while MAMBA's integration of local dynamics modeling with CTDE can mismatch global transition dynamics in Dec-POMDPs. Compared with IPPO under the same decentralized policy-learning setting, MODAL achieves faster convergence and stronger performance, highlighting the benefit of history-conditioned objective inference and weighted MF communication for mitigating temporal objective aliasing. Consistent test-return trends are reported in Appendix E.1.

6.3 World Model Prediction Accuracy

MODAL achieves more accurate multi-step prediction. Following Zhang et al. (2025c), we evaluate recursive local observation prediction on MOSMAC `4t`. Table 1 shows that MODAL yields the lowest errors over both 4- and 8-step rollouts. The 8-step rollout exceeds the training prediction horizon, providing an out-of-distribution (OOD) test of temporal generalization. Replacing the diffusion model with a Gaussian MLP in MODAL-G markedly increases prediction error, underscoring the benefit of diffusion modeling for multimodal futures. MODAL-1, trained only for one-step prediction ($H = 1$), degrades more over longer rollouts,

Table 1: **World model prediction error.** Cumulative ℓ_1 local observation prediction error over 4- and 8-step rollouts on MOSMAC `4t`.

Method	4-step	8-step
MODAL	$\mathbf{0.386}_{\pm 0.011}$	$\mathbf{0.889}_{\pm 0.021}$
MODAL-G	$0.884_{\pm 0.041}$	$1.629_{\pm 0.053}$
MODAL-1	$0.429_{\pm 0.011}$	$1.016_{\pm 0.024}$
DIMA	$0.599_{\pm 0.059}$	$1.180_{\pm 0.141}$

Table 2: **History encoder ablation.** Values are reported as percentages (mean ± standard deviation over 8 seeds); we report test win rates for MOSMAC and test success rates for Flatland.

MOSMAC			Flatland		
Scenario	MODAL	w/o f_ψ	Scenario	MODAL	w/o f_ψ
`3t`	$\mathbf{89.91}_{\pm 9.07}$	$74.40_{\pm 11.30}$	`5_agents`	$82.89_{\pm 5.73}$	$\mathbf{84.10}_{\pm 5.38}$
`4t`	$\mathbf{96.27}_{\pm 3.31}$	$91.81_{\pm 6.99}$	`10_agents`	$\mathbf{78.86}_{\pm 4.56}$	$72.58_{\pm 10.73}$
`8t`	$\mathbf{97.44}_{\pm 2.71}$	$59.24_{\pm 48.91}$	`20_agents`	$\mathbf{32.46}_{\pm 14.49}$	$22.70_{\pm 10.50}$
`12t`	$\mathbf{77.66}_{\pm 9.81}$	$27.44_{\pm 32.45}$			

highlighting the value of multi-step future modeling for long-horizon foresight. DIMA, another diffusion-based world model, outperforms MODAL-G but remains less accurate and more variable than MODAL, indicating stronger long-range predictive fidelity and stability of MODAL. Additional MOSMAC results and evaluation details are provided in Appendix E.2.

6.4 Ablation Studies

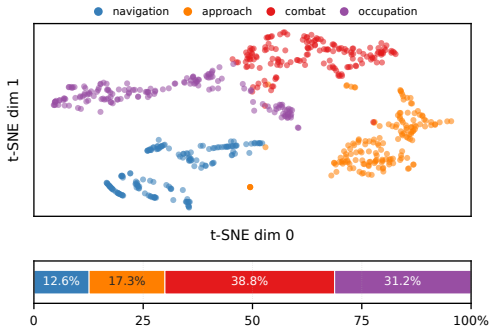


Figure 5: **History-conditioned representation visualization.** t-SNE shows z_t^i on MOSMAC `8t` across four event-defined stages; the bar indicates stage proportions.

History-Conditioned Objective Representation. Based on environment events, MOSMAC trajectories can be divided into four stages. Using the trained MODAL policy, we collect 100 trajectories from `8t`. Stage labels and sampling depend only on environment events, without using the learned representation or returns, avoiding circular interpretation. For visualization, we sample 160 representations per stage, while the stacked bar reports their original proportions. The t-SNE embedding in Figure 5 exhibits clear stage-dependent structure, indicating that the history encoder captures evolving objective context across distinct task stages. Consistently, Table 2 shows that removing the history encoder f_ψ substantially degrades performance and stability in MOSMAC `8t` and `12t`, with similar trends on larger Flatland settings. Together, these findings underscore the role of historical context in mitigating temporal objective aliasing by revealing task-stage distinctions obscured by instantaneous observations. Detailed stage definitions and sampling procedures are provided in Appendix D.

Weighted MF Communication. We evaluate weighted MF communication on the challenging MOSMAC `8t` and `12t` scenarios, where larger teams induce stronger interaction dependencies. We compare Weighted MF with three variants: *Unweighted MF + Stats* uses uniform aggregation with presence and density statistics, *Unweighted MF* removes these statistics, and *w/o MF* removes MF communication entirely. As shown in Figure 6, Weighted MF achieves the highest win rates in both scenarios. In `8t`, adding MF communication and neighborhood statistics progressively improves learning, while in `12t`, uniform aggregation degrades more substantially, indicating its limitation in capturing heterogeneous neighbor influence under stronger coupling. These results support combining distance-aware aggregation with explicit neighborhood statistics for reliable coordination. Corresponding test return results are provided in Appendix E.3.

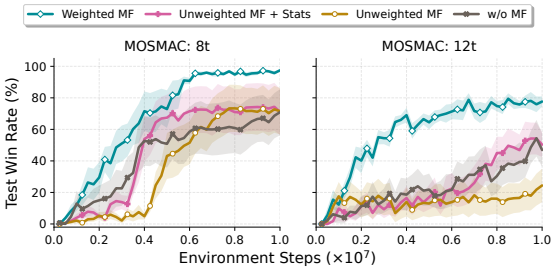


Figure 6: **Weighted MF ablation.** Curves show the mean test win rate over 8 random seeds, with shaded regions indicating the standard error.

7 Conclusion

We introduced **MODAL**, a decentralized diffusion world model for temporal objective aliasing in partially observable MARL. By combining history-conditioned representations with weighted MF communication, MODAL models multimodal local futures and supplies history-consistent imagined experience for policy learning. Our analysis bounds conditional trajectory and value-profile errors, while experiments on MOSMAC and Flatland demonstrate improved performance, prediction accuracy, and scalability in long-horizon coordination. These findings highlight history-conditioned generative modeling with efficient weighted MF communication as a promising direction for scalable decentralized coordination.

AI USE STATEMENT

In this work, we used generative AI tools as auxiliary support for language refinement and presentation, including grammar checking, wording suggestions, and improvements to the clarity and fluency of the manuscript. We also used these tools to assist with the readability and consistency of mathematical notation and technical descriptions, without introducing new theoretical results or scientific claims. Generative AI was not used to formulate the core research contributions, determine experimental outcomes, or derive the conclusions of this work. All AI-assisted content was carefully reviewed and verified by the authors for technical correctness and consistency with the underlying implementation and experiments. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on methodological research in multi-agent reinforcement learning and is evaluated entirely in simulated benchmarks, including MOSMAC and Flatland. The study involves no human participants and does not use personal, sensitive, or identifiable information. All training data are generated through agent-environment interactions within these simulated environments. The experiments follow established academic benchmarks and are conducted solely for research purposes.

REPRODUCIBILITY STATEMENT

We provide the complete online training procedure of MODAL in Algorithm 1, including online interaction, world model learning, imagined rollout, and policy optimization. Complete proofs of the theoretical results, including the weighted MF approximation and trajectory and value error analyses, are provided in Appendix B. Detailed environment configurations, computational resources, and key implementation and training hyperparameters are reported in Appendix C and Appendix D. The code for training and evaluation is available at `https://anonymous.4open.science/r/MODAL-7906`.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-1873.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. TransDreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022. URL https://arxiv.org/abs/2202.09481.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 4735–4763, 2023a. URL https://proceedings.mlr.press/v202/chen23q.html.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023b. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Kai Cui, Christian Fabian, Anam Tahir, and Heinz Koeppl. Major-minor mean field multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 9603–9632, 2024. URL https://proceedings.mlr.press/v235/cui24a.html.

Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the StarCraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020. URL https://arxiv.org/abs/2011.09533.

Yuanrui Duan, Wengang Zhou, Yufeng Shi, Xiancheng Gao, Lin Liu, and Houqiang Li. Offline multi-agent reinforcement learning with global moderate generalization. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1986–1994, 2026. doi: 10.65109/CVXE7640.

Vladimir Egorov and Alexei Shpilman. Scalable multi-agent model-based reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 381–390, 2022. URL https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p381.pdf.

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

Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. MOSMAC: A multi-agent reinforcement learning benchmark on sequential multi-objective tasks. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 867–876, 2025. URL https://www.ifaamas.org/Proceedings/aamas2025/pdfs/p867.pdf.

Matthias Gerstgrasser and David C Parkes. Meta-RL for multi-agent RL: Learning to adapt to evolving agents. In *Meta-Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://openreview.net/forum?id=0toY1f8-Iq9.

Raj Ghugare, Homanga Bharadhwaj, Benjamin Eysenbach, Sergey Levine, and Russ Salakhutdinov. Simplifying model-based RL: Learning representations, latent-space models, and policies with one objective. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=MQcmfgRxf7a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Qianyue Hao, Wenzhen Huang, Tao Feng, Jian Yuan, and Yong Li. GAT-MF: Graph attention mean field for very large scale multi-agent reinforcement learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 685–697, 2023. doi: 10.1145/3580305.3599359.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Shengchao Hu, Li Shen, Ya Zhang, and Dacheng Tao. Learning multi-agent communication from graph modeling perspective. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/37c6d0bc4d2917dcbea693b18504bd87-Abstract-Conference.html.

Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. UPDeT: Universal multi-agent RL via policy decoupling with transformers. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=v9c7hr9ADKx.

Tianmeng Hu, Biao Luo, Chunhua Yang, and Tingwen Huang. MO-MIX: Multi-objective multi-agent cooperative decision-making with deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10):12098–12112, 2023. doi: 10.1109/TPAMI.2023.3283537.

Zican Hu, Zongzhang Zhang, Huaxiong Li, Chunlin Chen, Hongyu Ding, and Zhi Wang. Attention-guided contrastive role representations for multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/66e226469f20625aaebddbe47f0ca997-Abstract-Conference.html.

Jia-Hua Lee, Bor-Jiun Lin, Wei-Fang Sun, and Chun-Yi Lee. EDELINE: Enhancing memory in diffusion-based world models via linear-time sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-3526.

Xinran Li, Xiaolu Wang, Chenjia Bai, and Jun Zhang. Exponential topology-enabled scalable communication in multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3514dbacaebf0f38b25adfe59ed81a8a-Abstract-Conference.html.

Zhuhui Li, Chunbo Luo, Liming Huang, Luyu Qi, and Geyong Min. Achieving equilibrium under utility heterogeneity: An agent-attention framework for multi-agent multi-objective reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 29538–29545, 2026. doi: 10.1609/aaai.v40i35.40196. Issue 35.

Zongyue Li, Xiao Han, Yusong Li, Niklas Strauß, and Matthias Schubert. DAWM: Diffusion action world models for offline reinforcement learning via action-inferred transitions. In *Workshop on Building Physically Plausible World Models at the 42nd International Conference on Machine Learning (ICML)*, 2025b. URL https://arxiv.org/abs/2509.19538.

Qihan Liu, Jianing Ye, Xiaoteng Ma, Jun Yang, Bin Liang, and Chongjie Zhang. Efficient multi-agent reinforcement learning by planning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d74e6bfe9ce029526e69db14d2c281ec-Abstract-Conference.html.

Yat Long Richie Lo, Biswa Sengupta, Jakob Foerster, and Michael Noukhovitch. Learning multi-agent communication with contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=vZZ4hhniJU.

Chengdong Ma, Aming Li, Yali Du, Hao Dong, and Yaodong Yang. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence*, 6:1006–1020, 2024. doi: 10.1038/s42256-024-00879-7.

Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Alexandre Lacoste, and Sai Rajeswar. Choreographer: Learning and adapting skills in imagination. In *Offline Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://openreview.net/forum?id=BxYsP-7ggf.

Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, Gereon Vienken, Irene Sturm, Guillaume Sartoretti, and Giacomo Spigler. Flatland-RL: Multi-agent reinforcement learning on trains. *arXiv preprint arXiv:2012.05893*, 2020. URL https://arxiv.org/abs/2012.05893.

Hyungho Na and Il-Chul Moon. LAGMA: LAtent goal-guided multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 37122–37140, 2024. URL https://proceedings.mlr.press/v235/na24b.html.

Hyungho Na, Kwanghyeon Lee, Sumin Lee, and Il-Chul Moon. Trajectory-class-aware multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9a7358e7778a9c1facbfe53cb26e459-Abstract-Conference.html.

Siddharth Nayak, Adelmo Morrison Orozco, Marina Ten Have, Vittal Thirumalai, Jackson Zhang, Darren Chen, Aditya Kapoor, Eric Robinson, Karthik Gopalakrishnan, James Harrison, Brian Ichter, Anuj Mahajan, and Hamsa Balakrishnan. Long-horizon planning for multi-agent robots in partially observable environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-2169.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning (ICML)*, volume 70, pp. 2681–2690, 2017. URL https://proceedings.mlr.press/v70/omidshafiei17a.html.

Georgios Papoudakis, Filippos Christianos, and Stefano V. Albrecht. Agent modelling under partial observability for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 19210–19222, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/a03caec56cd82478bf197475b48c05f9-Abstract.html.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021b. URL https://openreview.net/forum?id=cIrPX-Sn5n.

Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=9CcgO0LhKG.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=TdBaDGCpjly.

David Rother, Joni Pajarinen, Jan Peters, and Thomas H. Weisswange. Open-ended coordination for multi-agent systems using modular open policies. *Autonomous Agents and Multi-Agent Systems*, 39:40, 2025. doi: 10.1007/s10458-025-09723-7.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 2186–2188, 2019. URL https://www.ifaamas.org/Proceedings/aamas2019/pdfs/p2186.pdf. Extended Abstract.

Amutheezan Sivagnanam, Ava Pettet, Hunter Lee, Ayan Mukhopadhyay, Abhishek Dubey, and Aron Laszka. Multi-agent reinforcement learning with hierarchical coordination for emergency responder stationing. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 45813–45834, 2024. URL https://proceedings.mlr.press/v235/sivagnanam24a.html.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, volume 37, pp. 2256–2265, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Haolin Song, Mingxiao Feng, Wengang Zhou, and Houqiang Li. MA2CL: Masked attentive contrastive learning for multi-agent reinforcement learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4226–4234, 2023. doi: 10.24963/ijcai.2023/470. URL https://www.ijcai.org/proceedings/2023/470.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Taisuke Takayama, Naoto Yoshida, and Tadahiro Taniguchi. Dreamer-CPC: Message learning with world models for decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:2607.19809*, 2026. URL https://arxiv.org/abs/2607.19809.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Fan Wang, Zhiyuan Chen, Yuxuan Zhong, Sunjian Zheng, Pengtao Shao, Bo Yu, Shaoshan Liu, Jianan Wang, Ning Ding, Yang Cao, and Yu Kang. Context and diversity matter: The emergence of in-context learning in world models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a46b9dcef23e9160f46d800a49a931d0-Abstract-Conference.html.

Han Wang, Binbin Chen, Tieying Zhang, and Baoxiang Wang. Learning to communicate through implicit communication channels. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/89b89c04f55ea7c7ca989992bb6a98c0-Abstract-Conference.html.

Di Xue, Jing Jiang, Shaowei Zhang, Wenhao Guo, Lei Yuan, Zongzhang Zhang, and Yang Yu. Learning disentangled multi-agent world model for decentralized control. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/61713.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 80, pp. 5571–5580, 2018. URL https://proceedings.mlr.press/v80/yang18d.html.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24611–24624, 2022. doi: 10.52202/068431-1787.

Lei Yuan, Lihe Li, Ziqian Zhang, Fuxiang Zhang, Cong Guan, and Yang Yu. Multi-agent continual coordination via progressive task contextualization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 36(4):6326–6340, 2025. doi: 10.1109/TNNLS.2024.3394513.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. COMBO: Compositional world models for embodied multi-agent cooperation. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7d03c6bf9f07acb4038eea96c63db52d-Abstract-Conference.html.

Weinan Zhang, Xihuai Wang, Jian Shen, and Ming Zhou. Model-based multi-agent policy optimization with adaptive opponent-wise rollouts. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3384–3391, 2021. doi: 10.24963/ijcai.2021/466.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. STORM: Efficient stochastic transformer based world models for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-1182.

Yang Zhang, Chenjia Bai, Bin Zhao, Junchi Yan, Xiu Li, and Xuelong Li. Decentralized transformers with centralized aggregation are sample-efficient multi-agent world models. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=xT8BEgXmVc.

Yang Zhang, Xinran Li, Jianing Ye, Shuang Qiu, Delin Qu, Xiu Li, Chongjie Zhang, and Chenjia Bai. Revisiting multi-agent world modeling from a diffusion-inspired perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025c. doi: 10.52202/085713-2194.

Yi Zhao, Aidan Scannell, Yuxin Hou, Tianyu Cui, Le Chen, Dieter Büchler, Arno Solin, Juho Kannala, and Joni Pajarinen. Generalist world model pre-training for efficient reinforcement learning. In *World Models Workshop at International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/37543.

Zijie Zhao, Honglei Guo, Shengqian Chen, Kaixuan Xu, Bo Jiang, Yuanheng Zhu, and Dongbin Zhao. Empowering multi-robot cooperation via sequential world models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=IvUM6UwYCJ.

Chongyi Zheng, Seohong Park, Sergey Levine, and Benjamin Eysenbach. Intention-conditioned flow occupancy models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10007910.

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 (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-0136.

APPENDIX CONTENTS

A Training Procedure of MODAL **16**

B Proofs for the Theoretical Analysis **16**

B.1 Weighted MF Approximation 16

B.2 Proof of Theorem 1 18

B.3 Proof of Corollary 1 19

B.4 Proof of Theorem 2 19

C Environment Details **19**

D Implementation Details **21**

E Additional Experimental Results **22**

E.1 Additional Policy Evaluation 22

E.2 Additional World Model Prediction Evaluation 23

E.3 Additional Weighted MF Ablation 24

F Limitations and Future Work **25**

A TRAINING PROCEDURE OF MODAL

This section summarizes the complete online training procedure of MODAL in Algorithm 1, including online interaction, world model learning, imagined rollout, and policy optimization.

B PROOFS FOR THE THEORETICAL ANALYSIS

B.1 WEIGHTED MF APPROXIMATION

To complete the error characterization in Eq. (17), we analyze the approximation induced by compressing the complete visible-neighbor observations into the distance-weighted MF representation. Throughout the theoretical analysis, we neglect network fitting error to isolate the effects of history representation and weighted MF aggregation.

For agent i with a nonempty neighborhood, MODAL directly aggregates the observations of its visible neighbors as

$$\bar{o}_t^i = \sum\nolimits_{j\in\mathcal{C}_t^i} w_{ij} o_t^j, \qquad \sum\nolimits_{j\in\mathcal{C}_t^i} w_{ij} = 1. \tag{21}$$

The presence indicator h_t^i and neighborhood density ρ_t^i are additional empirical components introduced in MODAL to retain neighborhood presence and cardinality information. Since both are computed exactly from $\mathcal{C}_t^i$, they introduce no additional weighted-mean approximation error and are treated as fixed conditioning variables below.

Omitting common conditioning variables for brevity, let $\boldsymbol{\epsilon}^*(z, \mathbf{o}_t^i)$ denote the oracle conditional noise predictor given the complete neighborhood observations $\mathbf{o}_t^i$. Following the weighted pairwise decomposition of Hao et al. (2023), we assume its neighborhood-dependent component admits

$$\boldsymbol{\epsilon}^*(z, \mathbf{o}_t^i) = \sum\nolimits_{j\in\mathcal{C}_t^i} w_{ij} \boldsymbol{\epsilon}(z, o_t^j). \tag{22}$$

For each neighbor j, we express its observation as the weighted mean plus a local fluctuation, $o_t^j = \bar{o}_t^i + \delta o_t^{ij}$, where $\delta o_t^{ij} := o_t^j - \bar{o}_t^i$. Expanding each term in Eq. (22) around $\bar{o}_t^i$ with a second-order

Algorithm 1 Online training procedure of MODAL

Require: History length L and rollout horizon H

Initialize online replay buffer $\mathcal{D}$, history encoder f_ψ, diffusion model $\boldsymbol{\epsilon}_\theta$, auxiliary heads, policy π_ϕ, critic V_ν, and target critic $V_{\bar{\nu}}$.
for *each training epoch* **do**
  ONLINE INTERACTION
  for *each environment timestep* t **do**
    for *each agent* i **do**
      Construct m_t^i using Eq. (9).
      Compute z_t^i using Eq. (10).
      Form $\boldsymbol{\xi}_t^i = [o_t^i, m_t^i, z_t^i]$ and sample $a_t^i \sim \pi_\phi(\cdot \mid \boldsymbol{\xi}_t^i)$.
    end for
    Execute the joint action, observe the next observations, utilities, termination signals, and available-action masks, and store the transition in $\mathcal{D}$.
  end for
  WORLD MODEL LEARNING
  Sample length-$(L+H)$ trajectory segments from $\mathcal{D}$.
  for *each sampled agent trajectory* **do**
    Compute z_t^i using Eq. (10) and construct $\mathbf{c}_t^i = [o_t^i, a_t^i, m_t^i, z_t^i]$.
    Construct $\mathbf{x}_t^i(0) = [o_{t+1}^i, \ldots, o_{t+H}^i]$ and sample its noisy version $\mathbf{x}_t^i(\tau)$ using Eq. (3).
    Update the history encoder f_ψ, diffusion model $\boldsymbol{\epsilon}_\theta$, and auxiliary heads using Eq. (14).
  end for
  IMAGINED ROLLOUT
  Sample rollout starting contexts from $\mathcal{D}$ and initialize the imagined histories with the corresponding real histories.
  for $h = 0, \ldots, H-1$ **do**
    for *each agent* i **do**
      Construct $\hat{m}_{t+h}^i$ using Eq. (9).
      Compute $\hat{z}_{t+h}^i$ from the latest length-L imagined interaction history using Eq. (10).
      Form $\hat{\boldsymbol{\xi}}_{t+h}^i = [\hat{o}_{t+h}^i, \hat{m}_{t+h}^i, \hat{z}_{t+h}^i]$ and sample $\hat{a}_{t+h}^i \sim \pi_\phi(\cdot \mid \hat{\boldsymbol{\xi}}_{t+h}^i)$.
      Sample the H-step observation trajectory by iteratively denoising with the diffusion model $\boldsymbol{\epsilon}_\theta$ and retain the first predicted observation as $\hat{o}_{t+h+1}^i$.
      Sample the corresponding $\hat{u}_{t+h}^i$, $\hat{d}_{t+h}^i$, and $\hat{\mathbf{v}}_{t+h}^i$ with the auxiliary heads.
      Store the imagined transition for policy optimization.
    end for
    Stop trajectories whose predicted termination signal is active.
  end for
  POLICY OPTIMIZATION
  Update policy π_ϕ over imagined trajectories using Eq. (15).
  Update critic V_ν over imagined trajectories using Eq. (16), and periodically update $V_{\bar{\nu}}$.
end for

Taylor approximation gives

$$\begin{aligned}\boldsymbol{\epsilon}^*(z, \mathbf{o}_t^i) &= \sum\nolimits_{j \in \mathcal{C}_t^i} w_{ij} \left[\boldsymbol{\epsilon}(z, \bar{o}_t^i) + \nabla_o \boldsymbol{\epsilon}(z, \bar{o}_t^i)\, \delta o_t^{ij} + R_t^{ij} \right] \\ &= \boldsymbol{\epsilon}(z, \bar{o}_t^i) + \nabla_o \boldsymbol{\epsilon}(z, \bar{o}_t^i) \sum\nolimits_{j \in \mathcal{C}_t^i} w_{ij} \delta o_t^{ij} + \sum\nolimits_{j \in \mathcal{C}_t^i} w_{ij} R_t^{ij} \\ &= \boldsymbol{\epsilon}(z, \bar{o}_t^i) + \sum\nolimits_{j \in \mathcal{C}_t^i} w_{ij} R_t^{ij}, \end{aligned} \tag{23}$$

where R_t^{ij} denotes the second-order Taylor remainder. The first-order term vanishes because the fluctuations satisfy $\sum_{j \in \mathcal{C}_t^i} w_{ij} \delta o_t^{ij} = 0$ by Eq. (21).

We assume that the local neighborhood dispersion around the weighted MF observation is uniformly bounded as

$$\left\| o_t^j - \bar{o}_t^i \right\|_2^2 = \left\| o_t^j \right\|_2^2 + \left\| \bar{o}_t^i \right\|_2^2 - 2 (o_t^j)^\top \bar{o}_t^i \le 2C, \qquad j \in \mathcal{C}_t^i, \tag{24}$$

where C is a finite constant determined by the magnitude and dispersion of the local observation vectors.

Under the weighted pairwise decomposition in Eq. (22), assume that $\boldsymbol{\epsilon}$ has an $L_m(\tau)$-Lipschitz Jacobian with respect to the neighbor observation. Following Yang et al. (2018), the second-order Taylor remainder satisfies $\|R_t^{ij}\|_2 \leq \frac{L_m(\tau)}{2}\|\delta o_t^{ij}\|_2^2$. Taken together with Eqs. (23) and (24) and $\sum_j w_{ij} = 1$, this gives

$$\left\|\boldsymbol{\epsilon}^*(z, \mathbf{o}_t^i) - \boldsymbol{\epsilon}(z, \bar{o}_t^i)\right\|_2 \leq \sum_{j \in \mathcal{C}_t^i} w_{ij}\|R_t^{ij}\|_2 \leq \frac{L_m(\tau)}{2}\sum_{j \in \mathcal{C}_t^i} w_{ij}\|o_t^j - \bar{o}_t^i\|_2^2 \leq L_m(\tau)C. \tag{25}$$

Thus, replacing the complete neighbor observations with their weighted MF representation incurs an approximation error bounded by $L_m(\tau)C$.

B.2 Proof of Theorem 1

Proof. For the VP-SDE, the forward diffusion process is

$$d\mathbf{x}(\tau) = -\frac{1}{2}\beta(\tau)\mathbf{x}(\tau)\,d\tau + \sqrt{\beta(\tau)}\,d\mathbf{w}(\tau), \qquad \bar{\alpha}(\tau) = \exp\left(-\int_0^\tau \beta(s)\,ds\right), \tag{26}$$

where $\mathbf{w}(\tau)$ denotes a standard Wiener process. This process yields a perturbation variance of $1 - \bar{\alpha}(\tau)$, with the score and noise parameterizations satisfying

$$\nabla_{\mathbf{x}(\tau)} \log p_\tau(\mathbf{x}(\tau) \mid \mathbf{c}) = -\frac{\boldsymbol{\epsilon}(\mathbf{x}(\tau), \tau, \mathbf{c})}{\sqrt{1 - \bar{\alpha}(\tau)}}. \tag{27}$$

Accordingly, the true and MODAL reverse processes have drifts

$$\begin{aligned} b^{i,*}(\tau) &= -\frac{1}{2}\beta(\tau)\mathbf{x}(\tau) + \frac{\beta(\tau)}{\sqrt{1 - \bar{\alpha}(\tau)}}\boldsymbol{\epsilon}^*\left(\mathbf{x}(\tau), \tau, z_t^{i,*}, \mathbf{o}_t^i\right), \\ \widetilde{b}^i(\tau) &= -\frac{1}{2}\beta(\tau)\mathbf{x}(\tau) + \frac{\beta(\tau)}{\sqrt{1 - \bar{\alpha}(\tau)}}\boldsymbol{\epsilon}\left(\mathbf{x}(\tau), \tau, z_t^i, m_t^i\right). \end{aligned} \tag{28}$$

Suppressing their common diffusion-state arguments, Eq. (17) therefore gives

$$\left\|b^{i,*}(\tau) - \widetilde{b}^i(\tau)\right\|_2 \leq \frac{\beta(\tau)}{\sqrt{1 - \bar{\alpha}(\tau)}}\delta_\epsilon(\tau). \tag{29}$$

The two reverse processes share the diffusion coefficient $\sqrt{\beta(\tau)}\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix of the same dimension as $\mathbf{x}(\tau)$. Following Chen et al. (2023b), applying Girsanov's theorem to the reverse diffusion paths and the data-processing inequality to the terminal output $\mathbf{x}(0)$, i.e., the generated H-step local trajectory, yields

$$D_{\mathrm{KL}}\left(\mathbb{P}_H^{i,\boldsymbol{\pi}}\middle\|\widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}\right) \leq \Delta_T + \frac{1}{2}\int_0^T \mathbb{E}\left[\frac{\left\|b^{i,*}(\tau) - \widetilde{b}^i(\tau)\right\|_2^2}{\beta(\tau)}\right] d\tau, \tag{30}$$

where Δ_T denotes the finite-time mismatch between the terminal forward distribution and the Gaussian prior at diffusion time T.

Substituting Eq. (29) into Eq. (30) gives

$$D_{\mathrm{KL}}\left(\mathbb{P}_H^{i,\boldsymbol{\pi}}\middle\|\widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}\right) \leq \Delta_T + \frac{1}{2}\int_0^T \frac{\beta(\tau)}{1 - \bar{\alpha}(\tau)}\delta_\epsilon^2(\tau)\,d\tau := \Xi, \tag{31}$$

which establishes Eq. (18).

Finally, applying Pinsker's inequality (Tsybakov, 2009, Lemma 2.5) gives

$$D_{\mathrm{TV}}\left(\mathbb{P}_H^{i,\boldsymbol{\pi}}, \widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}\right) \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_H^{i,\boldsymbol{\pi}}\middle\|\widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}\right)} \leq \sqrt{\frac{\Xi}{2}} := \delta_P. \tag{32}$$

□

As the terminal forward distribution approaches the Gaussian prior used to initialize the reverse process, $\Delta_T \to 0$. The remaining trajectory distribution error is therefore governed by the history representation and weighted MF approximation errors contained in $\delta_\epsilon(\tau)$. Since $1 - \bar{\alpha}(\tau) \to 0$ in the low-noise limit, the uniform error bound does not directly guarantee integrability at $\tau = 0$. Following the early-stopping treatment of Chen et al. (2023a), the practical reverse diffusion is terminated at a sufficiently small positive diffusion time, avoiding the singular zero-noise endpoint.

B.3 Proof of Corollary 1

Proof. Fix $\mathcal{E}_t^i$ and $a \in \mathcal{A}^i$. For the local trajectory $\mathbf{x}_t^i = [o_{t+1}^i, \ldots, o_{t+H}^i]$, define

$$F_{t,H}^{i,\boldsymbol{\pi}}(x) := \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{H-1} \gamma^k u_{t+k}^i \,\middle|\, \mathbf{x}_t^i = x, \mathcal{E}_t^i, a_t^i = a\right]. \tag{33}$$

By the law of total expectation, the true and model-induced continuation values are, respectively,

$$Q_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i, a) = \mathbb{E}_{\mathbb{P}_H^{i,\boldsymbol{\pi}}}\left[F_{t,H}^{i,\boldsymbol{\pi}}(\mathbf{x}_t^i)\right], \tag{34}$$

$$\widetilde{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i, a) := \mathbb{E}_{\widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}}\left[F_{t,H}^{i,\boldsymbol{\pi}}(\mathbf{x}_t^i)\right], \tag{35}$$

where the latter evaluates the same utility functional under MODAL's trajectory distribution.

Under the bounded-utility assumption $|u_t^i| \le U_{\max}$,

$$\left|F_{t,H}^{i,\boldsymbol{\pi}}(x)\right| \le \sum_{k=0}^{H-1} \gamma^k U_{\max} = U_{\max}\frac{1-\gamma^H}{1-\gamma} := U_H. \tag{36}$$

Following Ma et al. (2024), we apply the total-variation argument, under which, for arbitrary probability distributions $\mathbb{P}$ and $\mathbb{Q}$ and any bounded measurable function f, the difference in expectations satisfies

$$|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]| \le 2\|f\|_\infty D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}). \tag{37}$$

Applying Eq. (37) to $f = F_{t,H}^{i,\boldsymbol{\pi}}$, together with Eq. (36) and Theorem 1, gives

$$\left|Q_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i, a) - \widetilde{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i, a)\right| \le 2U_H D_{\mathrm{TV}}\left(\mathbb{P}_H^{i,\boldsymbol{\pi}}, \widetilde{\mathbb{P}}_H^{i,\boldsymbol{\pi}}\right) \le 2U_H\delta_P, \tag{38}$$

which proves Eq. (19). □

B.4 Proof of Theorem 2

Proof. For brevity, let $\mathbf{q}_t = \mathbf{Q}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i)$ and $\widetilde{\mathbf{q}}_t = \widetilde{\mathbf{Q}}_H^{i,\boldsymbol{\pi}}(\mathcal{E}_t^i)$, and define their counterparts at t' analogously. For histories satisfying Eq. (2), the reverse triangle inequality and Corollary 1 give

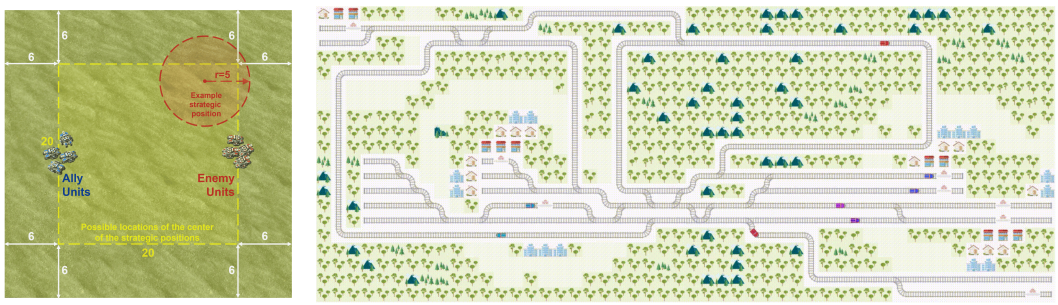
$$\begin{aligned}\|\widetilde{\mathbf{q}}_t - \widetilde{\mathbf{q}}_{t'}\|_\infty &\ge \|\mathbf{q}_t - \mathbf{q}_{t'}\|_\infty - \|\mathbf{q}_t - \widetilde{\mathbf{q}}_t\|_\infty - \|\mathbf{q}_{t'} - \widetilde{\mathbf{q}}_{t'}\|_\infty \\ &\ge \epsilon_Q - 4U_H\delta_P.\end{aligned} \tag{39}$$

Therefore, if $4U_H\delta_P < \epsilon_Q$, the model-induced value profiles remain distinct despite $\|o_t^i - o_{t'}^i\|_2 \le \epsilon_o$. Thus, sufficiently accurate diffusion world modeling preserves the action-conditioned continuation distinction hidden by instantaneous observations, providing a guarantee for MODAL to mitigate temporal objective aliasing.

□

C Environment Details

We evaluate MODAL on two cooperative multi-agent benchmarks with complementary long-horizon coordination challenges: MOSMAC (Geng et al., 2025) and Flatland (Mohanty et al., 2020). Representative instances of the two environments are illustrated in Figure 7. Detailed task and environment configurations used in our experiments are described below.



(a) MOSMAC (b) Flatland

Figure 7: Representative instances of the evaluation environments. MOSMAC requires agents to coordinate navigation toward a strategic position with combat against guarding enemies, whereas Flatland requires trains to coordinate long-horizon routes over shared railway infrastructure while avoiding congestion and deadlocks.

Table 3: MOSMAC single-task scenarios used in our experiments. All scenarios use siege tanks and plain terrain. "Move feat." denotes the dimension of movement-related observation features.

Setting	`3t`	`4t`	`8t`	`12t`
Allies	3	4	8	12
Enemies	3	4	8	12
Map size	32×32	32×32	32×32	32×32
Terrain	Plain	Plain	Plain	Plain
Move feat.	9	9	9	9
Timestep limit	50	50	100	100

MOSMAC. MOSMAC (Geng et al., 2025) extends the StarCraft Multi-Agent Challenge (Samvelyan et al., 2019; Ellis et al., 2023) to multi-objective cooperative decision-making and formulates the problem as a multi-objective decentralized partially observable Markov decision process. The benchmark considers combat, safety, and navigation objectives and contains both single-task and sequential multi-task scenarios. We evaluate MODAL on all four single-task siege-tank scenarios, namely `3t`, `4t`, `8t`, and `12t`. These scenarios preserve the same task structure while progressively increasing the number of interacting agents.

In each episode, allied units must cooperatively reach a strategic position (SP) while enemy units guard the target and engage approaching agents. At the beginning of an episode, the two coordinates of the SP are independently sampled from $\mathcal{U}(6, 26)$ on the 32×32 map, preventing agents from exploiting a fixed navigation target. Each agent can move in four directions, attack an enemy unit, or take a no-op action, while its local observation contains movement feasibility together with the direction and distance to the SP. An episode terminates if (i) the allied team is defeated; (ii) at least one allied unit survives, all enemies have been eliminated, and every surviving allied unit is within a radius of 5 from the strategic position; or (iii) the timestep limit is reached.

For the selected scenarios, the combat and navigation objectives are linearly scalarized with equal weights $(0.5, 0.5)$. We use dense reward signals consisting of combat-progress and navigation-progress terms, together with an enemy-death bonus of 10 and a terminal win bonus of 200. Reward scaling is enabled with a scaling rate of 20. The scenario specifications are summarized in Table 3.

Flatland. Flatland (Mohanty et al., 2020) models railway rescheduling as cooperative multi-agent routing on a two-dimensional rail network. Each train is assigned a starting location and a target destination, while its motion is constrained by the topology and directionality of the railway tracks. A train selects among five discrete actions: no-op, move left, move forward, move right, and stop, subject to the available rail transitions. Since trains share tracks and junctions, their interactions can lead to congestion, blocking, and deadlocks, requiring coordinated decisions over extended tempo-

Table 4: Flatland configurations used in our experiments, based on the released MAMBA implementation (Egorov & Shpilman, 2022). "City grid" denotes `grid_distribution_of_cities`.

Setting	`5_agents`	`10_agents`	`20_agents`
Grid size	35×35	35×35	40×60
Trains	5	10	20
Cities	3	4	6
City grid	False	False	False
Inter-city rails	2	2	2
In-city rails	4	4	4
Malfunction rate	$1/100$	$1/150$	$1/200$

ral horizons. Stochastic train malfunctions further introduce uncertainty by temporarily preventing affected trains from proceeding.

We use three Flatland configurations with 5, 10, and 20 agents, respectively, based on the environment settings of MAMBA (Egorov & Shpilman, 2022). We note a discrepancy between the MAMBA paper and its released implementation for the largest scenario: the paper reports a 15-agent configuration, whereas the released code actually instantiates 20 agents on a 40×60 grid with 6 cities. To avoid ambiguity, we refer to this actual instantiated setting as `20_agents` throughout this paper and report its corresponding environment parameters.

These environments progressively increase the number of trains and cities, while the largest configuration also increases the rail-network size, yielding denser traffic and more challenging coordination. In our experiments, the shared reward combines a $+10$ arrival reward, distance-progress shaping with coefficient 0.01, and a one-time -5 penalty upon deadlock. The environment configurations are summarized in Table 4.

D Implementation Details

MODAL is implemented on top of PyMARL using PyTorch. All experiments were conducted on a machine equipped with eight NVIDIA GeForce RTX 4090 GPUs and an AMD EPYC 7763 64-core processor, with each run using a single GPU. For baseline methods, we followed their official implementations and adopted the recommended hyperparameters to ensure fair comparisons.

Network architecture. Since all evaluated environments consist of homogeneous agents, we adopt parameter sharing across agents to improve sample efficiency and reduce training cost while preserving decentralized execution. The main architectural settings are summarized in Table 5. Following Hafner et al. (2025), observations are transformed element-wise using $\mathrm{symlog}(x) = \mathrm{sign}(x)\log(1 + |x|)$, which compresses large-magnitude values while retaining finer resolution around zero. Utilities are represented as two-hot categorical targets over bins obtained by applying symexp, the inverse of symlog, to an evenly spaced grid in the symlog domain. The utility head is trained with categorical cross-entropy, and scalar predictions are decoded as the probability-weighted mean of the bin values. We use 51 bins for MOSMAC and the `5_agents` Flatland scenario, and 61 bins for the `10_agents` and `20_agents` Flatland scenarios.

Model optimization. The history encoder and diffusion world model are jointly optimized under the world model objective, allowing the historical representation and trajectory predictor to be learned consistently from the same model-based training signal. The diffusion model $\boldsymbol{\epsilon}_\theta$ is optimized with the standard noise-prediction objective, while the policy π_ϕ and critic V_ν are optimized with PPO-based objectives. The main optimization and training settings are summarized in Table 6.

MOSMAC stages and t-SNE sampling. For Figure 5, we collect 100 trajectories using the fully trained MODAL policy on MOSMAC `8t`. Stage labels are defined solely by environment events, independent of the learned history representation, returns, or t-SNE embedding. Specifically, navigation ends once the allied team continuously detects the enemy; approach continues until the first

Table 5: Architecture details of MODAL.

Hyperparameter	Value
History Encoder (f_ψ)	
Architecture	GRU + linear projection
History context length L	8 in MOSMAC 16 in Flatland
GRU hidden dimension	32 in MOSMAC 128 in Flatland
Context embedding dimension	16 in MOSMAC 64 in Flatland
Normalization	LayerNorm
Diffusion Model (ϵ_θ)	
Backbone	Temporal 1-D U-Net
Local observation trajectory horizon H	4 in MOSMAC 8 in Flatland
Base channel dimension	128
Channel multipliers	$[1, 4, 8]$
Residual blocks per resolution	2
Activation	SiLU
Diffusion-time embedding dimension	32
Training diffusion timesteps	64
DDIM sampling steps	16
Variance schedule	Cosine
Cosine schedule offset	0.008
Maximum variance	0.999
Auxiliary Heads	
Architecture	3-layer MLP
Hidden dimension	128
Activation	ReLU
Policy (π_ϕ)	
Architecture	Linear-GRU-Linear
GRU hidden dimension	64
Critic (V_ν)	
Architecture	3-layer MLP
Hidden dimension	64

enemy casualty or sustained coordinated attacks involving at least 35% of surviving allies, marking the onset of combat. Occupation begins either when fewer than 50% of the enemies remain and the team approaches the strategic-position region, or when all enemies have been eliminated. For balanced t-SNE visualization, we uniformly sample 160 representations from each stage, while the stacked bar preserves their original timestep proportions across the collected trajectories.

E Additional Experimental Results

E.1 Additional Policy Evaluation

Figure 8 reports test returns on MOSMAC and Flatland. Consistent with the main results, MODAL remains competitive on the simpler `3t`, `4t`, and `5_agents` scenarios, while its advantage grows with coordination complexity. On MOSMAC, MODAL converges faster and achieves higher returns in `8t` and `12t`. On Flatland, it achieves the highest final returns in `10_agents` and `20_agents`, with a larger advantage in the latter despite strong early performance from MAPPO. Overall, these results further demonstrate MODAL's effectiveness for long-horizon multi-agent coordination as interaction dependencies increase.

Table 6: Optimization and training hyperparameters for MODAL.

Hyperparameter	**Value**
World model optimizer	Adam
Actor and critic optimizer	Adam
World model learning rate	3×10^{-4}
Actor and critic learning rate	3×10^{-4}
Utility loss weight λ_u	1.0
Termination loss weight λ_d	0.2
Action availability loss weight λ_a	0.2
Gradient-norm clipping	10
Discount factor γ	0.99
PPO clipping coefficient	0.1
PPO epochs per update	4
Target-critic update coefficient	0.01
Distance-decay rate κ	2.0 in MOSMAC 0.05 in Flatland
Entropy coefficient $\lambda_{\mathcal{H}}$	0.02 for MOSMAC `3t` and `4t` 0.01 for MOSMAC `8t` and Flatland 0.10 for MOSMAC `12t`
Parallel environments	10
Episode batch size	10
Replay buffer size	10 episodes

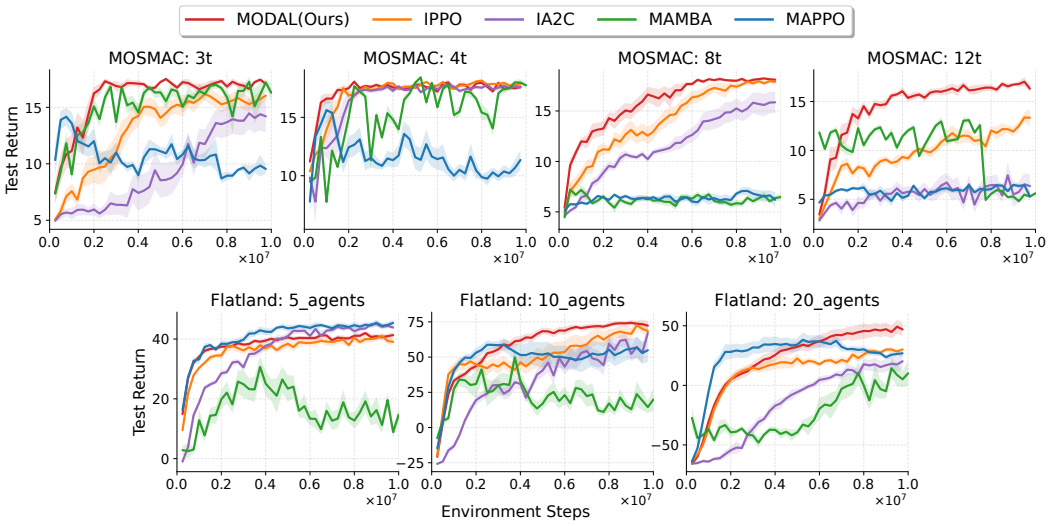


Figure 8: **Average test episode returns of all methods on MOSMAC and Flatland.** Curves show the mean performance over 8 random seeds, with shaded regions indicating the standard error.

E.2 Additional World Model Prediction Evaluation

We further evaluate local observation prediction accuracy on additional MOSMAC scenarios. For each scenario, we collect 20 real trajectories under a fixed environment seed by uniformly sampling from available actions, and construct 100 test segments, each containing 8 steps of ground-truth history followed by up to 8 future steps. All methods and training seeds are evaluated on the same test segments and ground-truth action sequences. Starting from the ground-truth history, predicted observations are recursively fed back for subsequent prediction without teacher forcing. We compute the mean absolute error (ℓ_1) over agents and observation features and accumulate it over 4 and 8 rollout steps. To examine the contribution of the diffusion trajectory predictor, MODAL-

Table 7: **Additional world model prediction errors on MOSMAC.** Cumulative ℓ_1 local observation error over 4- and 8-step rollouts. Values are mean $\pm$ standard deviation over 8 seeds.

	3t		8t	
Method	4-step	8-step	4-step	8-step
MODAL	$\mathbf{0.406}_{\pm \mathbf{0.019}}$	$\mathbf{0.911}_{\pm \mathbf{0.033}}$	$\mathbf{0.412}_{\pm \mathbf{0.014}}$	$\mathbf{0.877}_{\pm \mathbf{0.040}}$
MODAL-G	$0.825_{\pm 0.033}$	$1.553_{\pm 0.058}$	$0.697_{\pm 0.029}$	$1.302_{\pm 0.043}$
MODAL-1	$0.439_{\pm 0.028}$	$1.014_{\pm 0.063}$	$0.435_{\pm 0.051}$	$0.982_{\pm 0.162}$
DIMA	$0.579_{\pm 0.087}$	$1.170_{\pm 0.232}$	$0.585_{\pm 0.052}$	$1.125_{\pm 0.102}$

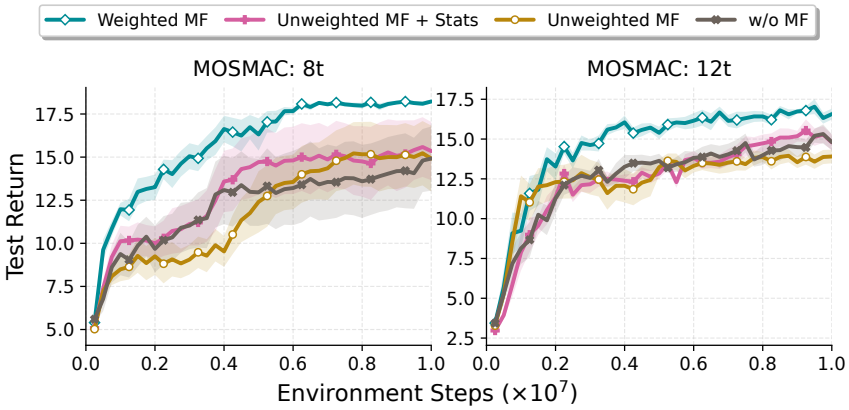


Figure 9: **Weighted MF ablation on MOSMAC `8t` and `12t` using test return.** Curves show the mean performance over 8 random seeds, with shaded regions indicating the standard error.

G replaces MODAL's diffusion-based local observation trajectory model with a 4-layer MLP that predicts the mean and standard deviation of Gaussian distributions over future local observation trajectories. MODAL-1 instead retains the diffusion-based world model backbone but predicts only the next local observation with prediction horizon $H = 1$. DIMA serves as a diffusion-based world model baseline, modeling global-state transitions conditioned on recent global states and joint actions through sequential denoising. For unified comparison, its predicted global states are decoded into agent-level local observations through its learned global-state-to-joint-observation decoder. Since all methods use 4-step imagined rollouts during policy learning, the 8-step evaluation extends beyond the training-time rollout length, providing an out-of-distribution (OOD) test of temporal generalization. As shown in Table 7, MODAL consistently achieves the lowest prediction error across both scenarios and rollout lengths while maintaining low variability across seeds. MODAL-G exhibits a clear degradation, supporting the benefit of diffusion-based trajectory modeling over Gaussian prediction for multimodal local futures. MODAL-1 remains competitive at shorter rollouts but deteriorates more noticeably as the rollout extends, indicating that explicitly modeling multi-step futures improves long-range temporal prediction and OOD generalization. DIMA performs better than MODAL-G yet remains less accurate and generally more variable than MODAL, further demonstrating the superior predictive fidelity and stability of the proposed diffusion-based world modeling approach.

E.3 Additional Weighted MF Ablation

Figure 9 reports the corresponding test return comparison on MOSMAC `8t` and `12t`. Consistent with the win rate results in Section 6.4, Weighted MF maintains clearly higher returns throughout the training and converges to the best final performance in both scenarios. The gap is especially evident in `8t`, where the simplified variants plateau at substantially lower returns, while `12t` shows a persistent advantage of the full design. These results further support combining distance-weighted aggregation with neighbor-presence and density statistics for robust interaction modeling in challenging multi-agent coordination tasks.

F LIMITATIONS AND FUTURE WORK

MODAL follows an online world model-based reinforcement learning paradigm, where the world model, policy, and value networks are alternately learned from scratch. Since imagined trajectories support policy and value learning, inaccurate world model predictions early in training can compound over imagined rollouts and destabilize optimization. Consequently, diffusion world model hyperparameters require careful tuning to maintain stable training. More robust model warm-up or uncertainty-aware imagination may alleviate this sensitivity. In addition, the current history encoder and auxiliary prediction heads adopt relatively simple architectures, as our focus is to validate the effectiveness of the proposed diffusion world model in mitigating temporal objective aliasing and improving long-horizon multi-agent learning with evolving objectives. More expressive temporal representations and auxiliary modeling may further improve its capability and scalability. Finally, our theoretical analysis encounters an endpoint singularity in the diffusion weighting term near zero diffusion time. We adopt an early-stopping treatment to keep the error bound well-defined, while future work could explore endpoint-aware analyses or tighter characterizations of score approximation errors to establish more complete theoretical guarantees over the full diffusion trajectory.