DAET: DISTRIBUTIONAL ACTION EQUIVALENCE TESTING INTEGRATING ALEATORIC AND EPISTEMIC UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern reinforcement learning (RL) algorithms typically treat actions as distinct during learning, even when different actions lead to nearly identical long-term consequences. This action redundancy unnecessarily inflates the effective exploration space and severely degrades exploration efficiency. This motivates the need for a principled criterion for identifying actions that can be safely treated as equivalent. However, existing equivalence criteria are inadequate for this purpose: transition-based equivalence is overly restrictive, while mean-based equivalence is too permissive and insensitive to risk. More importantly, both paradigms treat action equivalence as a property of directly accessible ground truth, whereas in online RL, it must be inferred through learned estimates from finite data. To bridge this gap, we introduce Distributional Action Equivalence Testing (DAET), a novel formulation that casts action equivalence as a statistical inference problem over return distributions. DAET jointly accounts for aleatoric uncertainty in long-term outcomes and epistemic uncertainty in their learned estimates, requiring sufficient statistical evidence before equivalence is accepted. Building on this formulation, we propose DAET-Mask, an uncertainty-aware framework that operationalizes DAET for online RL. It integrates a distributional RL backbone with Deep Ensembles, a dual-gating mechanism, and an uncertainty-guided ROPE decision rule to realize safe dynamic action masking without relying on hand-crafted rules. Theoretically, we show that distributional action masking is near-optimal in the ideal case and remains robust under bounded estimation error. Experiments in both controlled and realistic environments demonstrate that DAET-Mask reliably identifies and masks redundant actions, thereby improving exploration efficiency.

1 INTRODUCTION

Reinforcement learning (RL) has achieved remarkable success in complex decision-making tasks, but its efficiency is often constrained by the size and structure of the action space (Dulac-Arnold et al., 2015; Mnih et al., 2015; Silver et al., 2016; Berner et al., 2019; Vinyals et al., 2019). In many practical environments, different actions can lead to nearly identical long-term consequences, giving rise to the problem of action equivalence in MDPs (Givan et al., 2003; Li et al., 2006; Baram et al., 2021). For example, in robotic navigation with discretized steering commands, nearby angles such as 23° and 24° may move the agent to the same state, leading to nearly identical future trajectories. However, standard RL algorithms typically treat these actions as distinct, requiring the agent to repeatedly explore redundant choices. This action redundancy unnecessarily inflates the effective exploration space and severely degrades exploration efficiency. Therefore, it is critical to identify and abstract equivalent actions for efficient RL.

Existing approaches for identifying action equivalence in RL mainly fall into two paradigms. The first is transition-based equivalence (Ferns et al., 2004; Ravindran, 2004; Baram et al., 2021; Zhong et al., 2024), which defines two actions as equivalent only if they induce the same immediate rewards and transition dynamics. However, this criterion is overly restrictive: two actions may differ in one-step dynamics while leading to nearly identical long-term returns. Besides, in model-free RL, the required transition probabilities are typically unavailable. The second is mean-based equivalence (Even-Dar et al., 2006; Abel et al., 2016; Rosenfeld et al., 2017), which defines two actions

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.

as equivalent when their expected returns are equal. However, this criterion is too permissive. For example, one action may yield a deterministic return of 0, while another yields -10 or 10 with equal probability. Although the two actions have the same expectation, they clearly imply different risk profiles and should not be treated as equivalent. This makes the criterion risk-blind by relying solely on expectations. More importantly, both paradigms treat action equivalence as a property of directly accessible ground truth, whereas in online RL, it must be inferred from learned estimates based on finite data. In this context, assessing action equivalence is complicated by two sources of uncertainty (Kendall & Gal, 2017; Abdar et al., 2021). The first is aleatoric uncertainty, arising from the intrinsic stochasticity of long-term returns induced by stochastic environment dynamics. The second is epistemic uncertainty, arising from the agent's limited knowledge and imperfect estimation. Consequently, ignoring epistemic uncertainty can lead to unreliable equivalence assessments based on learned estimates. These limitations motivate a new criterion for action equivalence that integrates both the aleatoric uncertainty in the environment and the epistemic uncertainty in the agent.

To bridge this gap, we propose **D**istributional **A**ction **E**quivalence **T**esting (**DAET**), a new formulation that casts action equivalence as a hypothesis testing problem. Under DAET, two actions are considered practically equivalent only when sufficient statistical evidence supports the similarity of their long-term return distributions. This formulation addresses both sources of uncertainty simultaneously. First, it accounts for aleatoric uncertainty by defining equivalence in terms of full return distributions, thereby capturing the intrinsic variability and risk of long-term outcomes. Second, it accounts for epistemic uncertainty by using an epistemic confidence gate to ensure reliable estimates and formulating equivalence assessment as a hypothesis testing problem.

Building upon this testing formulation, we propose **D**istributional **A**ction **E**quivalence **T**esting with Dynamic **Mask**ing (**DAET-Mask**), an uncertainty-aware framework that operationalizes DAET in online RL. First, DAET-Mask employs a distributional RL backbone to account for aleatoric uncertainty in long-term outcomes. Second, it uses Deep Ensembles to quantify epistemic uncertainty in agent's estimates. Third, it performs equivalence testing using an uncertainty-guided Region of Practical Equivalence (ROPE) rule based on Wasserstein distance. Moreover, a dual-gating mechanism filters out action pairs that are clearly non-equivalent or insufficiently reliable before testing. Once an action is identified as redundant, it is temporarily masked for a finite validity horizon. After the horizon expires, the action becomes available again and must pass a new equivalence test before being masked. In this way, DAET-Mask dynamically narrows the action space to reduce inefficient exploration while preserving decision quality. Our contributions are summarized as follows.

- We introduce Distributional Action Equivalence Testing (DAET), a novel formulation that casts action equivalence in RL as a hypothesis testing problem, jointly accounting for aleatoric uncertainty in long-term outcomes and epistemic uncertainty in their learned estimates.
- We propose DAET-Mask, an uncertainty-aware framework that operationalizes DAET for online RL. It integrates a distributional RL backbone with Deep Ensembles, a dual-gating mechanism and an uncertainty-guided ROPE decision rule to realize safe dynamic action masking.
- We establish theoretical guarantees showing that distributional action masking is near-optimal in the ideal case and remains robust under bounded estimation error.
- Experiments in both controlled and realistic environments demonstrate that DAET-Mask reliably identifies and masks redundant actions, thereby improving exploration efficiency.

2 Problem Formulation

2.1 Standard and Distributional Reinforcement Learning

MDP setup. We consider a standard Markov decision process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the transition probability distribution, $R(s, a)$ is the immediate reward, and $\gamma \in [0, 1)$ is the discount factor. For a given policy π, the discounted random return from a state-action pair (s, a) is $Z^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$, where the trajectory is generated under $s_0 = s$ and $a_0 = a$. The random return satisfies the distributional Bellman equation (Bellemare et al., 2017):

$$Z^\pi(s, a) \overset{D}{=} R(s, a) + \gamma Z^\pi(S', A'), \tag{1}$$

where $S' \sim P(\cdot \mid s, a)$, $A' \sim \pi(\cdot \mid S')$, and $\overset{D}{=}$ denotes equality in distribution.

Standard RL. In standard reinforcement learning, Eq. (1) is typically simplified by taking expectations on both sides. This yields the standard action-value function $Q^\pi(s,a) := \mathbb{E}[Z^\pi(s,a)]$, which satisfies the expectation-based Bellman equation:

$$Q^\pi(s,a) = \mathbb{E}\big[R(s,a) + \gamma Q^\pi(S', A')\big]. \tag{2}$$

The learning process then minimizes the temporal-difference (TD) error derived from Eq. (2):

$$\mathcal{L}_{\text{std}}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\Big[\big(Q_\theta(s,a) - \big(r + \gamma Q_{\theta^-}(s',a')\big)\big)^2\Big], \tag{3}$$

where θ^- denotes the target network, $\mathcal{D}$ is the replay buffer, and a' is the next action sampled from the target policy. This expectation-based formulation collapses the return distribution into a scalar value and therefore does not explicitly represent the variability of long-term outcomes.

Distributional RL. In contrast, distributional RL directly models the distributional identity in Eq. (1) without reducing the return to its expectation (Bellemare et al., 2017; Dabney et al., 2018). The learning objective is therefore formulated as a distributional discrepancy minimization:

$$\mathcal{L}_{\text{dist}}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\Big[d\big(Z_\theta(s,a),\, r + \gamma Z_{\theta^-}(s',a')\big)\Big], \tag{4}$$

where $d(\cdot,\cdot)$ denotes a statistical discrepancy metric (e.g., Wasserstein distance). By representing the full return distribution rather than only its expectation, distributional RL captures the intrinsic environmental stochasticity. It therefore provides a natural representation of aleatoric uncertainty, enabling the agent to distinguish actions with identical expected returns but different risk profiles.

2.2 Distributional Action Equivalence Testing

A central question in action abstraction is how to determine whether two actions should be treated as equivalent. Existing approaches primarily rely on two paradigms: transition-based equivalence (Definition 4) and mean-based equivalence (Definition 5). However, the former is overly restrictive, whereas the latter is too permissive and risk-blind. To resolve this dichotomy, we argue that action equivalence should be defined via the full return distribution.

Definition 1 (Distributional Action Equivalence)**.** *Two actions $a_1, a_2 \in \mathcal{A}$ are said to be distributionally equivalent in state s under policy π if*

$$Z^\pi(s,a_1) \overset{D}{=} Z^\pi(s,a_2), \tag{5}$$

where $\overset{D}{=}$ denotes equality in distribution.

Definition 1 captures equivalence across full return distributions, but exact equality in distribution is too strict for practical stochastic settings. Following the concept of the Region of Practical Equivalence (ROPE) in Bayesian statistics (Kruschke, 2018; Kruschke & Liddell, 2018), we relax it using the 1-Wasserstein distance, a widely used metric in RL (Villani et al., 2009; Bellemare et al., 2017).

Definition 2 (ϵ-Distributional Action Equivalence)**.** *Given a tolerance threshold $\epsilon \geq 0$, two actions $a_1, a_2 \in \mathcal{A}$ are said to be ϵ-distributionally equivalent in state s under policy π if the 1-Wasserstein distance between their return distributions is bounded by ϵ:*

$$W_1(Z^\pi(s,a_1), Z^\pi(s,a_2)) \leq \epsilon. \tag{6}$$

When $\epsilon = 0$, the definition reduces to Definition 1.

Definition 2 establishes a robust criterion for action equivalence that explicitly accounts for aleatoric uncertainty. However, like the transition- and mean-based criteria, it remains an idealized criterion: it assumes access to the unobservable true return distributions $Z^\pi(s,a)$. From a statistical perspective, this can be formalized by recognizing that the true return distribution $Z^\pi(s,a)$ is governed by an underlying data-generating mechanism. Let θ^* denote the unknown true environment parameter such that $Z^\pi(s,a) \equiv Z^\pi_{\theta^*}(s,a)$. For a given state-action pair (s,a_1,a_2) under policy π, let

$$\delta^\pi(s;a_1,a_2) := W_1(Z^\pi_{\theta^*}(s,a_1), Z^\pi_{\theta^*}(s,a_2)) \tag{7}$$

denote the true discrepancy between the two actions. Because θ^* is unobservable, $\delta^\pi(s;a_1,a_2)$ is a deterministic but unknown target value. In online RL, the agent only observes a finite interaction dataset $\mathcal{D}$, yielding an estimated parameter θ and an empirical discrepancy point-estimate:

$$\widehat{\delta}^\pi_\theta(s;a_1,a_2) := W_1(Z^\pi_\theta(s,a_1), Z^\pi_\theta(s,a_2))\,. \tag{8}$$

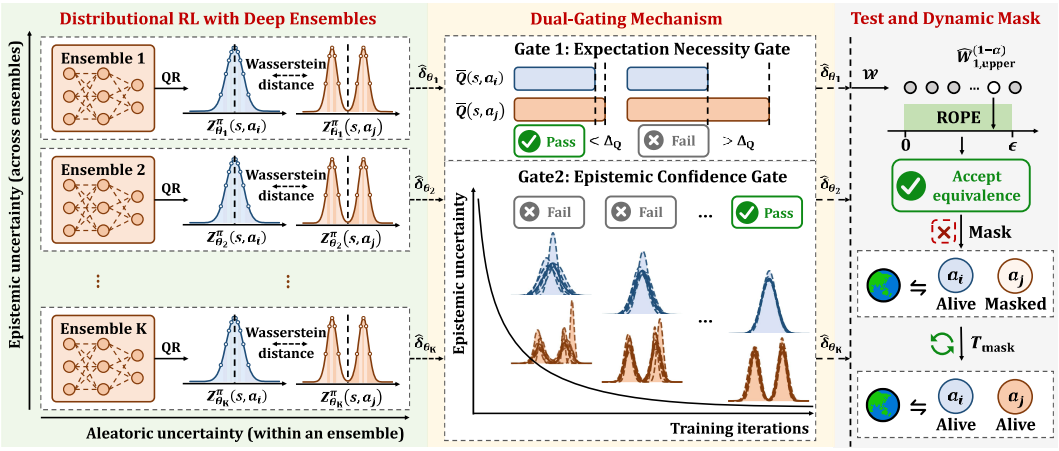


Figure 1: Overview of our proposed DAET-Mask.

However, directly evaluating equivalence via $\widehat{\delta}^{\pi}_{\theta}(s; a_1, a_2) \leq \epsilon$ is statistically fragile. Due to limited data and imperfect exploration, $\widehat{\delta}^{\pi}_{\theta}(s; a_1, a_2)$ is subject to epistemic uncertainty. Consequently, a small $\widehat{\delta}^{\pi}_{\theta}(s; a_1, a_2)$ may reflect genuine distributional equivalence or merely result from unreliable estimates. To address this fragility, we shift from direct point-estimate evaluation to rigorous statistical inference. Specifically, we formulate this hypothesis testing framework as follows:

Definition 3 (**D**istributional **A**ction **E**quivalence **T**esting (**DAET**))**.** *Given a state s, two actions $a_1, a_2 \in \mathcal{A}$, a policy π, and a tolerance threshold $\epsilon \geq 0$, action equivalence is formulated as the composite hypothesis test over the unknown true distributional discrepancy:*

$$\mathrm{H}_0 : \ \delta^{\pi}(s; a_1, a_2) \leq \epsilon \quad \textit{versus} \quad \mathrm{H}_1 : \ \delta^{\pi}(s; a_1, a_2) > \epsilon. \tag{9}$$

Equivalently, DAET tests whether the true parameter θ^* lies within the practical equivalence region:

$$\mathrm{H}_0 : \ \theta^* \in \Theta^{\pi}_{\epsilon}(s; a_1, a_2) \quad \textit{versus} \quad \mathrm{H}_1 : \ \theta^* \notin \Theta^{\pi}_{\epsilon}(s; a_1, a_2), \tag{10}$$

where $\Theta^{\pi}_{\epsilon}(s; a_1, a_2) := \{\theta : W_1\left(Z^{\pi}_{\theta}(s, a_1), Z^{\pi}_{\theta}(s, a_2)\right) \leq \epsilon\}$.

To operationalize this test, we adopt an uncertainty-aware learning procedure to construct an empirical distribution over plausible parameters θ from the observed data. It then induces a distribution over action discrepancies through the pushforward mapping $\theta \mapsto \delta^{\pi}_{\theta}(s; a_1, a_2)$. The resulting pushforward distribution over $\delta^{\pi}_{\theta}(s; a_1, a_2)$ provides statistical evidence about whether the unknown true discrepancy lies within the ROPE $[0, \epsilon]$. Specifically, let $\alpha \in (0, 1)$ denote the tolerance error level. DAET accepts the null hypothesis H_0 (i.e., asserts practical equivalence) if and only if the upper $1 - \alpha$ credible bound of the induced discrepancy distribution is no larger than ϵ:

$$\Pr\left(\delta^{\pi}_{\theta}(s; a_1, a_2) \leq \epsilon\right) \geq 1 - \alpha. \tag{11}$$

This testing perspective fundamentally distinguishes DAET from deterministic point-estimate comparisons. It ensures that practical equivalence is accepted if and only if the empirical discrepancy is not merely small, but reliably small under epistemic uncertainty. We emphasize that DAET is inherently a local, state-conditional formulation. A global notion of equivalence that requires actions to be indistinguishable across the entire state space is typically too restrictive to produce actionable abstraction in complex environments. A detailed discussion is provided in Appendix C.

3 The DAET-Mask Framework

To operationalize DAET in online RL, two main challenges must be addressed. First, standard RL architectures ignore the underlying aleatoric uncertainty and provide no quantification of epistemic uncertainty. Second, exhaustively executing rigorous hypothesis testing for all action pairs at every state is computationally prohibitive. To address these challenges, we propose **D**istributional **A**ction **E**quivalence **T**esting with Dynamic **Mask**ing (**DAET-Mask**). As illustrated in Figure 1, DAET-Mask builds on a distributional RL backbone augmented with Deep Ensembles to explicitly model

both aleatoric and epistemic uncertainties (Section 3.1). To overcome the computational bottleneck, a dual-gating mechanism efficiently screens out action pairs that are clearly non-equivalent or insufficiently reliable (Section 3.2). The remaining candidates are then rigorously evaluated by an uncertainty-guided ROPE decision rule (Section 3.3). Once an action is identified as redundant, it is temporarily masked and becomes available again upon expiration (Section 3.4). Finally, we establish theoretical safety guarantees for this distributional masking mechanism (Section 3.5).

3.1 Distributional RL with Deep Ensembles

To represent the long-term consequences of actions beyond their scalar expectations, we build on a distributional RL backbone. Specifically, we adopt the Quantile Regression (QR) parameterization (Dabney et al., 2018), which approximates the return distribution via a discrete uniform distribution over N learned quantile locations. To estimate epistemic uncertainty, we replace single-network prediction with Deep Ensembles (Lakshminarayanan et al., 2017). We train an ensemble of K independently initialized distributional models, parameterized by $\{\theta_k\}_{k=1}^{\mathrm{K}}$. Given a state-action pair (s, a), the ensemble yields K distinct return distribution estimates $\{Z_{\theta_k}(s, a)\}_{k=1}^{\mathrm{K}}$. Crucially, our framework is not restricted to value-based RL: in actor-critic architectures like PPO, the required distributional estimates can be generated by auxiliary quantile heads, leaving the original actor and scalar value head intact for standard policy optimization and advantage estimation.

This design disentangles two distinct sources of uncertainty. The dispersion within each predicted return distribution captures aleatoric uncertainty, i.e., intrinsic stochasticity in the environment. By contrast, disagreement across ensemble members captures epistemic uncertainty, i.e., uncertainty due to finite data or insufficient exploration. To summarize these predictions, we first extract the scalar expected return from each predicted distribution, denoted as $Q_k(s, a)$. We then compute the ensemble mean of these expected returns, denoted as $\bar{Q}(s, a)$. Finally, we quantify the epistemic uncertainty $\mathrm{U}_{\text{ep}}(s, a)$ using the coefficient of variation (CV) across the K predicted means:

$$\mathrm{U}_{\text{ep}}(s, a) = \frac{\sigma_Q(s, a)}{|\bar{Q}(s, a)| + c_{\text{stab}}}, \tag{12}$$

where $\sigma_Q(s, a)$ denotes the sample standard deviation of the ensemble predictions $\{Q_k(s, a)\}_{k=1}^{\mathrm{K}}$, and $c_{\text{stab}} > 0$ (e.g., 1) is a small constant for numerical stability.

3.2 The Dual-gating Mechanism

Performing rigorous distributional equivalence tests for all action pairs is both unnecessary and computationally prohibitive. Moreover, testing in regions of high epistemic uncertainty is statistically unreliable and may lead to premature masking. To address both issues, DAET-Mask employs a dual-gating mechanism to restrict testing to a small set of relevant and reliable candidate pairs. Given a state s, we first compute $\bar{Q}(s, a)$ and $\mathrm{U}_{\text{ep}}(s, a)$ for all $a \in \mathcal{A}$, and identify the current greedy action $a^* = \arg\max_{a \in \mathcal{A}} \bar{Q}(s, a)$. Each other action $a_i \neq a^*$ is then compared against a^* through two gates.

Gate 1: Expectation necessity gate. While mean-based equivalence is insufficient to define practical action redundancy, it provides a rigorous necessary condition for ϵ-distributional equivalence. We restrict the testing to actions whose expected returns are sufficiently close to the greedy action:

$$|\bar{Q}(s, a^*) - \bar{Q}(s, a_i)| \leq \Delta_Q. \tag{13}$$

By explicitly setting the threshold $\Delta_Q \geq \epsilon$, any action pair violating this condition trivially exceeds the equivalence threshold ϵ in Wasserstein distance. This transforms Gate 1 from a heuristic into a mathematically grounded pre-filter that safely reduces the search space without discarding any potentially equivalent actions. A formal derivation is provided in Appendix D.

Gate 2: Epistemic confidence gate. Rigorous testing should only proceed when the action estimates are sufficiently reliable. Thus, we require that epistemic uncertainty be bounded:

$$\max\left(\mathrm{U}_{\text{ep}}(s, a^*), \mathrm{U}_{\text{ep}}(s, a_i)\right) \leq \tau_{\text{ep}}, \tag{14}$$

where τ_{ep} is the epistemic tolerance threshold. If this gate is not passed, it implies that the agent has not adequately explored this state-action subspace, and testing is deferred to preserve exploration. Together, these two gates restrict full testing to action pairs that are both potentially ϵ-equivalent and sufficiently reliable for distributional evaluation.

3.3 Uncertainty-guided ROPE Decision Rule

To instantiate the theoretical testing criterion in Eq. (11), we leverage the ensemble of estimated return distributions to empirically approximate the pushforward discrepancy distribution. Specifically, for each candidate action pair (a^*, a_i) that passes the dual-gating mechanism, each ensemble head θ_k predicts the corresponding return distributions $Z^{\pi}_{\theta_k}(s, a^*)$ and $Z^{\pi}_{\theta_k}(s, a_i)$. Evaluating the discrepancy between these distributions across all K ensemble members yields:

$$\mathcal{W}(s; a^*, a_i) = \left\{ \widehat{\delta}^{\pi}_{\theta_1}(s; a^*, a_i), \ldots, \widehat{\delta}^{\pi}_{\theta_K}(s; a^*, a_i) \right\}. \tag{15}$$

Under the quantile parameterization, the 1-Wasserstein distance is efficiently computed as the mean absolute difference across aligned quantiles. Under this non-negative setting, the ROPE reduces to the one-sided interval $[0, \epsilon]$, and checking whether the Highest Density Interval (HDI) (Kruschke, 2014) of the discrepancy distribution is contained within the ROPE is mathematically equivalent to checking whether its upper bound does not exceed ϵ. Consequently, DAET-Mask implements a one-sided, quantile-based adaptation of the classical HDI–ROPE (Kruschke, 2018; Kruschke & Liddell, 2018) decision principle, declaring a_i practically equivalent to a^* if and only if

$$\widehat{W}^{(1-\alpha)}_{1,\text{upper}}(s; a^*, a_i) = \text{Quantile}\left(\mathcal{W}(s; a^*, a_i), 1-\alpha\right) \leq \epsilon, \tag{16}$$

where $\alpha \in (0, 1)$ is the tolerance error level defined in Eq. (11) (e.g., $\alpha = 0.05$ evaluates the empirical upper 0.95-quantile across ensembles). By using ensemble disagreement as a proxy for epistemic uncertainty, this mechanism guards against unreliable equivalence decisions and declares practical equivalence only when the estimated discrepancy provides sufficient evidence for equivalence.

3.4 Dynamic Action Masking

Building on the uncertainty-guided ROPE criterion, actions identified as practically equivalent to the current optimal action a^* are temporarily excluded from subsequent selections. At each decision state s, DAET-Mask constructs a state-conditioned mask $\mathrm{M}_t(s)$, with $\mathrm{M}_t(s, a) = 0$ for available actions and $\mathrm{M}_t(s, a) = -\infty$ for actions currently identified as redundant. For a logit-based policy, this mask directly regulates action selection via:

$$\pi_t(a|s) = \left[\text{Softmax}\left(\text{Logits}_t(s) + \mathrm{M}_t(s)\right)\right]_a, \tag{17}$$

By driving the selection probability of redundant actions to zero, the active mask prevents the policy from repeatedly allocating probability mass to actions that have been judged practically equivalent to the current greedy action, thereby redirecting exploration toward unresolved actions.

For discrete or frequently revisited states, we maintain a state-indexed dictionary to track the mask expiration. If an action a_i is identified as redundant at time t_0, its mask remains active until $t_0 + T_{\text{mask}}$. When the same state is revisited within the validity horizon, the stored mask is applied before action selection. Once the horizon expires, the action becomes available again and can be re-evaluated.

For continuous or high-dimensional state spaces, exact state recurrence is often rare, making persistent state-indexed storage inefficient. In this regime, DAET-Mask instead constructs a transient mask directly from the current state and applies it only to the current decision step, corresponding to $T_{\text{mask}} = 1$. Therefore, all candidate actions are reconsidered at the next decision event.

3.5 Theoretical Guarantee: Safe masking under Distributional Equivalence

In this section, we establish that the action space obtained by masking distributionally equivalent actions retains a policy whose value approximately preserves that of any fixed reference policy, with an optimal-control guarantee following as a special case. At any fixed interaction step, let $\mathcal{R}(s) \subseteq \mathcal{A}$ denote the set of currently masked actions at state s, $\varphi_s : \mathcal{R}(s) \to \mathcal{A} \backslash \mathcal{R}(s)$ map each masked action to a retained representative, and $\widehat{\Pi}$ denote the class of policies supported only on the retained actions.

Theorem 1 (Safe masking bound)**.** *Assume that return distributions have finite first moments. For any fixed policy π, suppose that every masked action has a retained representative satisfying*

$$W_1\left(Z^{\pi}(s, a), Z^{\pi}(s, \varphi_s(a))\right) \leq \epsilon, \qquad \forall s \in \mathcal{S},\ a \in \mathcal{R}(s). \tag{18}$$

Then, there exists a projected policy $\bar{\pi} \in \widehat{\Pi}$ such that

$$\left\| V^{\bar{\pi}} - V^{\pi} \right\|_{\infty} \leq \frac{\epsilon}{1-\gamma}. \tag{19}$$

The optimal-control guarantee follows directly by choosing the reference policy to be a fixed optimal policy $\pi^\star$ and defining $Z^\star(s,a) := Z^{\pi^\star}(s,a)$. Let $\widehat{V}^\star$ denote the optimal state-value function in the masked MDP. Because the projected policy $\bar{\pi}^\star$ is feasible in the masked MDP and masking restricts the feasible policy class, i.e., $V^{\bar{\pi}^\star}(s) \le \widehat{V}^\star(s) \le V^\star(s)$, Eq. (19) yields

$$0 \le V^\star(s) - \widehat{V}^\star(s) \le \frac{\epsilon}{1-\gamma}, \qquad \forall s \in \mathcal{S}. \tag{20}$$

In practice, equivalence testing relies on learned return-distribution estimates $Z^\pi_\theta(s,a)$. We extend our guarantee to accommodate heterogeneous local estimation errors.

Corollary 1 (Safe masking under estimation error)**.** *Suppose that for every action pair $(a, \varphi_s(a))$ accepted for masking, the estimated return distributions satisfy:*

(i) Local estimation accuracy: $W_1\left(Z^\pi(s,b), Z^\pi_\theta(s,b)\right) \le \eta(s,b)$ for each $b \in \{a, \varphi_s(a)\}$;

(ii) Estimated pairwise equivalence: $W_1\left(Z^\pi_\theta(s,a), Z^\pi_\theta(s,\varphi_s(a))\right) \le \epsilon$.

Then, the performance variation under the projected policy $\bar{\pi}$ satisfies

$$\left\|V^{\bar{\pi}} - V^\pi\right\|_\infty \le \frac{\epsilon_{\text{eff}}}{1-\gamma}, \quad \textit{where } \epsilon_{\text{eff}} = \sup_{s\in\mathcal{S}, a\in\mathcal{R}(s)} \left[\epsilon + \eta(s,a) + \eta(s,\varphi_s(a))\right]. \tag{21}$$

Corollary 1 only requires error control for the action pairs actually masked, requiring no uniform error bounds over unvisited or unmasked state-action regions. In DAET-Mask, the dual-gating mechanism filters out action pairs with large expected-return gaps or high epistemic uncertainty, restricting masking to well-controlled regimes. Furthermore, if there exists a uniform local error bound across all masked pairs $\eta(s,a) \le \eta$, the guarantee reduces to $(\epsilon + 2\eta)/(1-\gamma)$ as a conservative bound.

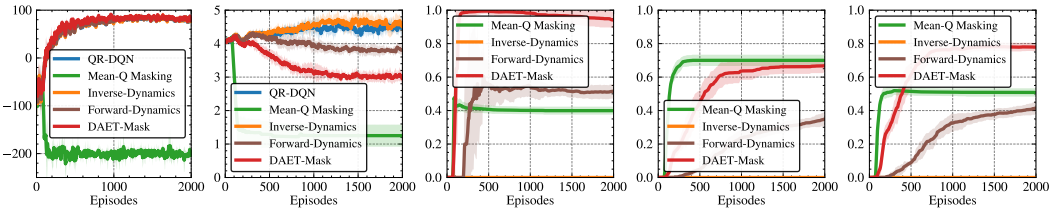
4 Experiments

Our empirical evaluation aims to answer two primary research questions: **RQ1:** Can DAET-Mask distinguish genuinely redundant actions and thereby achieve safe action masking? **RQ2:** Can the proposed dynamic masking mechanism translate into tangible efficiency improvements in complex, large-scale application scenarios? To answer these questions, we evaluate DAET-Mask in two complementary settings: a controlled GridWorld simulation designed to test the masking mechanism, and a realistic Robotic Mobile Fulfillment System (RMFS) task that assesses efficiency gains in complex scheduling. More experimental details are shown in Appendix I and J, respectively.

4.1 Simulation Study: GridWorld with Redundant and Risk-distinct Actions

Setup. To explicitly evaluate action masking mechanisms (**RQ1**), we construct a 10×10 GridWorld with a fork-join topology. Three obstacle walls partition the grid into regions connected by two cascaded fork-join gadgets. The action space comprises four standard directional moves, a deterministic duplicate, a risk-variant that induces the same transition and expected reward as right but with stochastic immediate reward, ten branch actions available exclusively at fork nodes, and a Merge action at branch states. We evaluate DAET-Mask against four baselines. The first is the standard QR-DQN, serving as the unmasked distributional RL reference (Dabney et al., 2018). The second is Mean-Q Masking, which masks actions solely by expected-return proximity (Li et al., 2006; Abel et al., 2016). The remaining two baselines probe the transition-based equivalence criterion through two learning-based variants: a Forward-Dynamics baseline that masks actions yielding identical predicted forward transitions (Abel et al., 2020; Gelada et al., 2019), and an Inverse-Dynamics baseline that identifies redundancy utilizing an inverse dynamics posterior (Baram et al., 2021; Zhong et al., 2024). Our evaluation comprises two axes: learning quality and masking correctness. Specifically, Episodic Return measures the cumulative reward per episode, while Action Space Size reports the average number of available actions per decision step, reflecting the effective reduction achieved by masking. Masking correctness is evaluated via standard precision, recall, and F1-score.

Results. As shown in Figure 2a, DAET-Mask achieves optimal performance comparable to QR-DQN, whereas Mean-Q Masking suffers a severe performance collapse. This collapse occurs because Mean-Q Masking aggressively reduces the action space to approximately one action per step



(a) Episodic Return (b) Action Space Size (c) Precision (d) Recall (e) F1-Score

Figure 2: Comparative evaluation of DAET-Mask and baseline methods in the fork-join GridWorld environment. Panels (a)–(e) show different metrics over training episodes, respectively. Solid lines and shaded regions represent the mean $\pm$ std across 10 random seeds.

(Figure 2b), mistakenly masking essential exploratory actions. In contrast, under the same threshold ($\Delta_Q = 0.3$), DAET-Mask safely reduces the action space from about 4.5 to 3.0 actions per step on average, improving exploration efficiency and leading to slightly faster convergence than the other methods. Meanwhile, the Forward-Dynamics and Inverse-Dynamics methods fail to achieve significant action space reduction due to their overly conservative criteria. In particular, Inverse-Dynamics fails to achieve any action space reduction, as its reliance on inverse posterior predictions struggles to learn discriminative redundancy representations in this highly symmetric topology. Moreover, all these baselines lack a mechanism to quantify epistemic uncertainty, and erroneous masking can easily occur early in training. To ensure a fair comparison and avoid early collapse, we explicitly delay the activation of all masking mechanisms until after 1,000 steps.

Figures 2c–2e further reveal the masking accuracy, with DAET-Mask achieving nearly perfect precision of 1.0 while maintaining a recall of 0.67. By securing high TPs while strictly bounding FPs near zero (Figure 4), DAET-Mask attains the highest F1-score (0.78). In contrast, Mean-Q Masking reaches a similar recall but generates substantially more FPs, leading to low precision and confirming its extensive over-masking behavior. On the other hand, the transition-based methods yield low recall and high FNs. Forward-Dynamics exhibits a considerable number of both FPs and FNs, while Inverse-Dynamics identifies almost no redundant actions, resulting in near-zero recall and F1-scores.

Overall, DAET-Mask achieves the best balance among action space reduction, masking correctness, and learning performance. Specifically, mean-based criteria determine action equivalence based on small differences in expected returns, making them risk-blind and prone to over-masking. Conversely, transition-based criteria require strictly identical one-step dynamics, making them unable to recognize multi-step distributional equivalences (e.g., branch actions), which leads to ineffective under-masking. DAET-Mask avoids both failure modes by using the expectation comparison only as a necessary-condition filter while making the equivalence assessment through the uncertainty-guided ROPE rule at the distributional level. These simulation results provide direct evidence for **RQ1**, confirming that DAET-Mask successfully executes safe and effective action masking.

4.2 Real-world Study: RMFS Order Allocation and Robot Scheduling

Setup. To evaluate the practical efficiency gains of DAET-Mask (**RQ2**), we consider a real-data Robotic Mobile Fulfillment System (RMFS) order allocation and robot scheduling task (Tang et al., 2026). RMFS is a goods-to-person warehouse system in which mobile robots transport inventory shelves between storage locations and workstations to fulfill dynamically arriving orders. The task is a real-time operations research (OR) problem with large, event-dependent action spaces: the agent must make sequential scheduling decisions under asynchronous warehouse events while supporting coupled order allocation and robot scheduling. We conduct experiments on real industrial warehouse data, including real warehouse layouts, inventory snapshots, and historical order streams.

The baselines are categorized into phased methods and joint methods according to their optimization structure. The phased methods combine three order allocation approaches (SQF (Liang et al., 2023), WLB (Tadumadze et al., 2023), and an OR-Tools solver) with four robot scheduling strategies (Nearest (Ma et al., 2017), Earliest (Deng et al., 2016), TSP (Li et al., 2024), and PSMDRL (Zhou et al., 2024)). For joint optimization, we include state-of-the-art methods including JOTP (Zhang

Table 1: Results on the real-world RMFS order allocation and robot scheduling task.

Type	Order Allocation	Robot Scheduling	Uses RL	Obj ↓	CompT ↓	Time ↓	ASR ↓
Phased	SQF	Nearest	No	618.62	259.91	0.60s	1.00
	SQF	Earliest	No	604.61	242.27	0.60s	1.00
	SQF	TSP	No	614.88	242.29	0.62s	–
	SQF	PSMDRL	Yes	550.59	218.33	2.50s	1.00
	WLB	Nearest	No	494.19	201.41	0.63s	1.00
	WLB	Earliest	No	606.33	257.18	0.73s	1.00
	WLB	TSP	No	520.75	204.66	0.89s	–
	WLB	PSMDRL	Yes	480.92	222.28	2.40s	1.00
	OR-Tools	Nearest	No	414.06	184.50	1m34s	–
	OR-Tools	Earliest	No	449.13	204.98	1m38s	–
	OR-Tools	TSP	No	416.62	185.66	1m39s	–
Joint	–	JOTP	Yes	528.69	205.28	0.95s	1.00
	–	SABS	No	465.05	192.74	1m50s	–
	–	SOAR	Yes	<u>406.79</u>	<u>152.36</u>	5.06s	<u>9.08%</u>
Ours	–	DAET-Mask	Yes	**394.10**	**145.65**	6.73s	**8.79%**
	–	Improvement	–	**3.1%**	**4.4%**	–	**3.2%**

Note. ASR is not applicable (–) for global or batch optimization algorithms (e.g., OR-Tools, TSP, SABS), as they evaluate complete solution spaces rather than operating within a step-by-step sequential action space.

et al., 2025), SABS (Yang et al., 2021), and SOAR (Tang et al., 2026). DAET-Mask first adopts SOAR's Top-K candidate filtering strategy to reduce the large candidate space, and further performs distributional equivalence testing to dynamically mask locally redundant actions. We evaluate task performance using three primary metrics: (1) Obj (Makespan), defined as the overall completion time of the scheduling process; (2) CompT, which measures the average duration from order arrival to fulfillment, reflecting service responsiveness; (3) Time, evaluating the computational time of each method per instance; and (4) ASR (Action Space Ratio), a pooled ratio over all decision events.

Results. As shown in Table 1, DAET-Mask achieves the best performance among the compared methods in both Obj and CompT. Specifically, compared with the strongest baseline SOAR, DAET-Mask reduces the makespan by 3.1% and the average order completion time by 4.4%. These performance gains are further reflected in the ASR metric. Beyond the Top-K candidate filtering used by SOAR, DAET-Mask further reduces the effective action space via distributional equivalence testing, decreasing ASR from 9.08% to 8.79%, a 3.2% relative reduction. This is particularly beneficial in the large, event-dependent action spaces of real-time warehouse operations, where asynchronous events and coupled order allocation and robot scheduling create a large number of possible scheduling decisions. By masking redundant actions, such as neighboring workstations with identical queuing implications or duplicate shelves with equivalent fulfillment potential, DAET-Mask reduces unnecessary exploration of actions that lead to effectively equivalent scheduling outcomes. Consequently, DAET-Mask improves exploration efficiency and allows the policy to focus its learning on genuinely distinct scheduling choices, contributing to the observed gains in scheduling performance.

Unlike the controlled GridWorld experiment, the RMFS task contains no manually injected duplicate actions. Instead, exploitable redundancy is implicit and emerges naturally from the operational structure of warehouse scheduling. Therefore, it evaluates DAET-Mask in a more realistic setting characterized by high-dimensional states, asynchronous events, and strongly coupled, multi-stage decisions. The simultaneous reduction in ASR and improvement in Obj and CompT suggest that DAET-Mask improves scheduling efficiency by reducing the effective action space. These results provide direct evidence for **RQ2**, showing that action-space reduction supports more efficient learning and translates into tangible performance gains in complex, large-scale OR applications.

5 Conclusion

We introduced DAET, a new formulation that redefines action equivalence in reinforcement learning under both aleatoric and epistemic uncertainty. Rather than comparing unobservable true quantities deterministically, DAET casts practical action equivalence as a hypothesis testing problem over return distributions. Building on this formulation, we proposed DAET-Mask to realize safe dynamic action masking. Both theoretical and experimental results demonstrate its efficiency.

AI USE STATEMENT

Large Language Models (LLMs) were used to assist with the writing and language editing of this manuscript. Specifically, LLMs were used for grammar checking, sentence rephrasing, improving readability, and refining the clarity and flow of the text. All research decisions, methodological developments, experimental designs, analyses, and conclusions were made and verified by the authors.

The authors take full responsibility for the content of this manuscript and have reviewed all LLM-assisted text to ensure its accuracy, originality, and consistency with the research.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923, New York, New York, USA, 2016. PMLR.

David Abel, Nate Umbanhowar, Khimya Khetarpal, Dilip Arumugam, Doina Precup, and Michael Littman. Value preserving state-action abstractions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1639–1650. PMLR, 2020.

Nir Baram, Guy Tennenholtz, and Shie Mannor. Action redundancy in reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 376–385. PMLR, 2021.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458, 2017.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Ondrej Biza and Robert Platt. Online abstraction with mdp homomorphisms for deep learning. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1125–1133. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning action representations for reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 941–950. PMLR, 2019.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Dingxiong Deng, Cyrus Shahabi, Ugur Demiryurek, and Linhong Zhu. Task selection in spatial crowdsourcing from worker's perspective. *GeoInformatica*, 20(3):529–568, 2016.

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. Fast reinforcement learning with large action sets using error-correcting output codes for mdp factorization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 180–194. Springer, 2012.

Gabriel Dulac-Arnold, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, and Ben Coppin. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

Eyal Even-Dar, Shie Mannor, Ece Mcgill Ca, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006. doi: 10.5555/1248547.1248586.

Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2170–2179. PMLR, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. In Roman Barták, Fazel Keshtkar, and Michael Franklin (eds.), *Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2022, Hutchinson Island, Jensen Beach, Florida, USA, May 15-18, 2022*, 2022. doi: 10.32473/flairs.v35i.130584. URL https://doi.org/10.32473/flairs.v35i.130584.

Shengyi Huang, Santiago Ontañón, Chris Bamford, and Lukasz Grela. Gym-μrts: Toward affordable full game real-time strategy games research with deep reinforcement learning. In *2021 IEEE Conference on Games (CoG)*, pp. 1–8. IEEE, 2021.

Anssi Kanervisto, Christian Scheller, and Ville Hautamäki. Action space shaping in deep reinforcement learning. In *2020 IEEE conference on games (CoG)*, pp. 479–486. IEEE, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

John K Kruschke and Torrin M Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic bulletin & review*, 25(1):178–206, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kunpeng Li, Tengbo Liu, PN Ram Kumar, and Xuefang Han. A reinforcement learning-based hyper-heuristic for agv task assignment and route planning in parts-to-picker warehouses. *Transportation research part E: logistics and transportation review*, 185:103518, 2024.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for mdps. In *International Symposium on Artificial Intelligence and Mathematics, AI&Math 2006, Fort Lauderdale, Florida, USA, January 4-6, 2006*, 2006. URL http://anytime.cs.umass.edu/aimath06/proceedings/P21.pdf.

Kaibo Liang, Li Zhou, Jianglong Yang, Huwei Liu, Yakun Li, Fengmei Jing, Man Shan, and Jin Yang. Research on a dynamic task update assignment strategy based on a "parts to picker" picking system. *Mathematics*, 11(7):1684, 2023.

Hang Ma, Jiaoyang Li, TK Kumar, and Sven Koenig. Lifelong multi-agent path finding for online pickup and delivery tasks. *arXiv preprint arXiv:1705.10868*, 2017.

Augustine N. Mavor-Parker, Matthew J. Sargent, Christian Pehle, Andrea Banino, Lewis D. Griffin, and Caswell Barry. Using forwards-backwards models to approximate mdp homomorphisms. *arXiv preprint arXiv:2209.06356*, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

Jason Pazis and Ronald Parr. Generalized value functions for large action sets. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1185–1192, 2011.

Balaraman Ravindran. *An Algebraic Approach to Abstraction in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2004.

Ariel Rosenfeld, Matthew E. Taylor, and Sarit Kraus. Speeding up tabular reinforcement learning using state-action similarities. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1722–1724. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Giorgi Tadumadze, Julia Wenzel, Simon Emde, Felix Weidinger, and Ralf Elbert. Assigning orders and pods to picking stations in a multi-level robotic mobile fulfillment system. *Flexible Services and Manufacturing Journal*, 35(4):1038–1075, 2023.

Yibang Tang, Yifan Yang, Jingyuan Wang, Junhua Chen, and Zhen Zhao. Soar: Real-time joint optimization of order allocation and robot scheduling in robotic mobile fulfillment systems. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 8065–8076, 2026.

Guy Tennenholtz and Shie Mannor. The natural language of actions. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6196–6205. PMLR, 2019.

Cédric Villani et al. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 2009.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Ziyi Wang, Xinran Li, Luoyang Sun, Haifeng Zhang, Hualin Liu, and Jun Wang. Learning state-specific action masks for reinforcement learning. *Algorithms*, 17(2):60, 2024.

Xiying Yang, Guowei Hua, Li Zhang, TC Cheng, and Tsan Ming Choi. Joint order assignment and picking station scheduling in kiva warehouses with multiple stations. *arXiv preprint arXiv:2108.09056*, 2021.

Tom Zahavy, Matan Haroush, Nadav Merlis, Daniel J. Mankowitz, and Shie Mannor. Learn what not to learn: Action elimination with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shaohui Zhang, Qiuying Han, Hai Zhu, Hongfeng Wang, Huiling Li, and Ke Wang. Real time task planning for order picking in intelligent logistics warehousing. *Scientific Reports*, 15(1):7331, 2025.

Dianyu Zhong, Yiqin Yang, and Qianchuan Zhao. No prior mask: Eliminate redundant action for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17078–17086, 2024.

Xuan Zhou, Xiang Shi, Wenqing Chu, Jingchen Jiang, Lele Zhang, and Fang Deng. Learning to solve multi-agv scheduling problem with pod repositioning optimization in rmfs. In *2024 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1–8. IEEE, 2024.

A CODE

The code is available at `https://anonymous.4open.science/r/daet-mask_public/`.

B THE WASSERSTEIN METRIC

To quantify the discrepancy between two probability distributions, we employ the Wasserstein metric, which is particularly well-suited for Distributional RL due to its stability with respect to disjoint supports. For two random variables U and V with cumulative distribution functions (CDFs) F_U and F_V, the p-Wasserstein distance is defined as:

$$W_p(U,V) = \left(\inf_{\pi \in \Pi(U,V)} \int_{\mathbb{R}\times\mathbb{R}} |u-v|^p \mathrm{d}\pi(u,v) \right)^{1/p}, \tag{22}$$

where $\Pi(U,V)$ denotes the set of all joint distributions (couplings) whose marginals are U and V.

In this work, we primarily focus on the 1-Wasserstein distance (W_1), which intuitively measures how much mass must be transported to transform one distribution into the other. Expressed as an expectation, the 1-Wasserstein distance takes the form:

$$W_1(U,V) = \inf_{\pi \in \Pi(U,V)} \mathbb{E}_{(u,v)\sim\pi}[\,|u-v|\,]. \tag{23}$$

In our setting, W_1 is particularly suitable for two reasons. First, unlike mean-based comparisons, it measures discrepancy between full return distributions and therefore reflects differences in variability and risk. Second, for one-dimensional return distributions represented by quantiles, it admits a computationally tractable closed-form solution via the inverse CDFs F_U^{-1} and F_V^{-1}:

$$W_1(U,V) = \int_0^1 |F_U^{-1}(\tau) - F_V^{-1}(\tau)| \mathrm{d}\tau. \tag{24}$$

This property allows for highly efficient distance computation when returns are parameterized by quantile regression networks (e.g., QR-DQN).

C THE GLOBAL NOTION OF DAET

In the main text, DAET is formulated as a local, state-conditioned notion of action equivalence. This choice reflects the fact that the decision consequences of an action are generally state-dependent: two actions may induce nearly indistinguishable long-term return distributions in one state, while leading to substantially different outcomes in another. For completeness, we discuss in this appendix a possible global counterpart of DAET and explain why it is not adopted as the primary abstraction mechanism in DAET-Mask.

A natural way to define global distributional action equivalence is to aggregate state-wise distributional discrepancies over a state visitation distribution. Let d_π denote the state visitation distribution induced by policy π, and define the local distributional discrepancy between two actions as

$$\delta_\pi(s; a_1, a_2) = W_1(Z^\pi(s,a_1), Z^\pi(s,a_2))\,. \tag{25}$$

Then we can define a visitation-weighted global action distance as

$$\mathrm{D}_\pi^{\text{avg}}(a_1,a_2) = \mathbb{E}_{s\sim d_\pi}\left[\delta_\pi(s;a_1,a_2)\right]. \tag{26}$$

Under this definition, two actions are globally ϵ-equivalent if

$$\mathrm{D}_\pi^{\text{avg}}(a_1,a_2) \le \epsilon, \tag{27}$$

or equivalently if the corresponding global hypothesis

$$\mathrm{H}_0^{\text{global}} : \mathrm{D}_\pi^{\text{avg}}(a_1,a_2) \le \epsilon \quad \text{versus} \quad \mathrm{H}_1^{\text{global}} : \mathrm{D}_\pi^{\text{avg}}(a_1,a_2) > \epsilon \tag{28}$$

is accepted. This formulation asks whether two actions are effectively equivalent over the states that the agent actually visits, rather than over the entire state space. It is therefore more realistic than

requiring equality for all $s \in \mathcal{S}$, especially in high-dimensional or continuous environments where exhaustive state-wise testing is infeasible.

However, an expectation-based global distance can suffer from a risk-dilution effect. If two actions are nearly identical in most frequently visited states but differ catastrophically in a small subset of states, the averaged distance $\mathrm{D}_{\pi}^{\mathrm{avg}}(a_1, a_2)$ may still be small. For example, suppose that two navigation actions have almost identical return distributions on flat terrain, but near a cliff one action remains safe while the other leads to failure. If cliff states are rare under d_π, their contribution may be diluted by the expectation, causing the global test to accept equivalence even though merging the actions would be unsafe in precisely the states where the distinction matters most. Thus, average global equivalence is not sufficient for safety-critical action abstraction.

A stricter alternative is to replace the expectation with a risk-sensitive functional over the random variable $\delta_\pi(s; a_1, a_2)$. For example, one may consider an upper-tail criterion such as

$$\mathrm{D}_{\pi,\beta}^{\mathrm{CVaR}}(a_1, a_2) = \mathrm{CVaR}_\beta\left(\delta_\pi(s; a_1, a_2)\right), \quad s \sim d_\pi, \tag{29}$$

where CVaR represents the Conditional Value at Risk, and CVaR_β measures the average discrepancy over the highest-discrepancy tail of visited states. The resulting global test

$$\mathrm{H}_0^{\mathrm{global}} : \mathrm{D}_{\pi,\beta}^{\mathrm{CVaR}}(a_1, a_2) \le \epsilon \tag{30}$$

is more robust to rare but high-impact state regions than the expectation-based criterion. In the limiting case, one may use a worst-case criterion,

$$\mathrm{D}_{\pi}^{\max}(a_1, a_2) = \sup_{s \in \mathrm{supp}(d_\pi)} \delta_\pi(s; a_1, a_2), \tag{31}$$

which accepts global equivalence only if the two actions are distributionally close throughout the support of the visitation distribution.

Although such risk-sensitive global criteria are conceptually appealing, they are typically too conservative for practical action masking. In realistic environments, truly global action equivalence is rare: actions are usually equivalent only in particular states or local regions of the state space. For instance, two routing choices may be interchangeable under normal traffic conditions but differ sharply when one route is blocked; two warehouse targets may be equivalent under one queue configuration but not under another. A global CVaR or worst-case criterion would reject equivalence as soon as a small subset of states exhibits meaningful divergence. As a result, the agent would be prevented from exploiting valid local redundancies in the majority of states where the actions are in fact interchangeable. This would make the abstraction safe but overly conservative, substantially limiting its ability to improve exploration efficiency.

Global DAET also introduces substantial statistical and computational difficulties. Estimating $\mathrm{D}_{\pi}^{\mathrm{avg}}$ or $\mathrm{D}_{\pi,\beta}^{\mathrm{CVaR}}$ requires collecting sufficiently representative samples from d_π, which itself changes throughout online learning. Moreover, evaluating global equivalence for all action pairs would require distributional comparisons across many visited states and numerous candidate pairs, leading to a cost that scales poorly with both the state coverage and the action-space size. Tail-sensitive criteria further increase the sample complexity, as reliable estimation of rare high-discrepancy regions requires substantially more data than estimating an average discrepancy.

For these reasons, DAET adopts a local, state-conditioned interpretation of distributional action equivalence. Instead of asking whether two actions are globally interchangeable, DAET asks whether a candidate action is practically equivalent to the current greedy action at the present state:

$$\mathrm{H}_0(s) : W_1(Z^\pi(s, a^*), Z^\pi(s, a_i)) \le \epsilon. \tag{32}$$

When this local hypothesis is accepted with sufficient epistemic confidence, DAET-Mask masks a_i only at state s. The action remains available in other states, where its consequences may differ from those of a^*. Thus, local masking avoids both failure modes of global abstraction: it does not average away dangerous local differences, and it does not reject useful local equivalences merely because the same actions differ elsewhere.

The global notion can therefore be viewed as a strict special case of the local formulation. If two actions satisfy a strong global criterion, such as

$$\sup_{s \in \mathrm{supp}(d_\pi)} W_1(Z^\pi(s, a_1), Z^\pi(s, a_2)) \le \epsilon, \tag{33}$$

then they also satisfy the local DAET criterion at every visited state. In this case, global action abstraction is justified. However, DAET-Mask does not require such a strong condition. By maintaining a state- and time-conditioned dynamic mask, it exploits equivalence wherever it is statistically supported, preserves action availability at other states, and restores a masked action once its validity horizon expires. This design is better aligned with the structure of online RL, where action redundancy is often local, policy-dependent, and revealed gradually through interaction.

D Mathematical Guarantee of Gate 1: The Expectation Necessity Gate

In Section 3.2, we introduce an expectation-based filtering gate (Gate 1) as a necessary condition for distributional equivalence. Here we provide the formal guarantee for this bound using the properties of the 1-Wasserstein distance.

Proposition 1. *Let U and V be two real-valued random variables with finite first moments. Then, the absolute difference between their expectations is bounded by their 1-Wasserstein distance:*

$$|\mathbb{E}[U] - \mathbb{E}[V]| \leq W_1(U, V). \tag{34}$$

Proof. By the Kantorovich-Rubinstein duality (Villani et al., 2009), the 1-Wasserstein distance can be expressed as:

$$W_1(U, V) = \sup_{\|f\|_L \leq 1} \left| \int_{\mathbb{R}} f(x) \mathrm{d}P_U(x) - \int_{\mathbb{R}} f(x) \mathrm{d}P_V(x) \right|, \tag{35}$$

where the supremum is taken over the class of all 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$. Because the supremum evaluates the maximum difference across this entire function class, substituting any specific 1-Lipschitz function naturally establishes a lower bound for $W_1(U, V)$. Consider the identity function $f(x) = x$, which is trivially 1-Lipschitz. Evaluating the dual representation with this specific function directly yields:

$$W_1(U, V) \geq \left| \int_{\mathbb{R}} x \mathrm{d}P_U(x) - \int_{\mathbb{R}} x \mathrm{d}P_V(x) \right| = |\mathbb{E}[U] - \mathbb{E}[V]|. \tag{36}$$

This completes the proof. □

Implication for DAET-Mask. In our framework, we test for ϵ-distributional equivalence, defined as $W_1(Z^\pi(s, a_1), Z^\pi(s, a_2)) \leq \epsilon$. The proposition above implies that for the true return distributions:

$$W_1(Z^\pi(s, a_1), Z^\pi(s, a_2)) \leq \epsilon \implies |Q^\pi(s, a_1) - Q^\pi(s, a_2)| \leq \epsilon. \tag{37}$$

Therefore, under the true expectations, any action pair with an expectation gap strictly greater than ϵ is guaranteed to violate the distributional equivalence hypothesis H_0. In practice, DAET-Mask relies on learned estimates $\bar{Q}$ rather than the unobservable true expectations Q^π. To safely account for estimation errors, Gate 1 sets the threshold $\Delta_Q \geq \epsilon$. By maintaining this margin, we ensure that no action pair that could potentially satisfy the ROPE decision rule is prematurely discarded, while efficiently screening out the vast majority of clearly sub-optimal actions.

E Discussion of Gate 2: The Epistemic Confidence Gate

In Section 3.2, we introduce Gate 2 as a confidence-based screening condition before full distributional equivalence testing. We quantify epistemic uncertainty at this stage by the disagreement of ensemble members in their predicted expected returns (Eq. (12)). This scalar quantity is used only for candidate filtering; the final equivalence decision remains fully distributional, since it is based on ensemble-wise Wasserstein distances and the ROPE decision rule.

We intentionally define the uncertainty score $\mathrm{U}_{\mathrm{ep}}(s, a)$ in Gate 2 using the disagreement of ensemble members solely over their predicted expected returns (Eq. (12)). While the epistemic uncertainty of a distributional prediction may in principle involve disagreement over the full set of predicted quantiles, using such a distribution-level metric already at the screening stage would significantly increase

computational cost. Instead, Gate 2 adopts a lightweight proxy that efficiently assesses whether an action estimate is sufficiently stable to merit further evaluation. Distribution-level epistemic uncertainty is then handled explicitly in the subsequent ROPE stage through ensemble-wise Wasserstein disagreement. Thus, distribution-level uncertainty is not ignored, but deferred to the final testing stage where it is most relevant.

F Proof of Theorem 1

F.1 Policy-relative Bound

Proof. For each state s, define the projected policy $\bar{\pi}$ by reallocating the probability mass of every masked action to its retained representative:

$$\bar{\pi}(b \mid s) = \begin{cases} 0, & b \in \mathcal{R}(s), \\ \pi(b \mid s) + \sum_{a \in \mathcal{R}(s):\varphi_s(a)=b} \pi(a \mid s), & b \notin \mathcal{R}(s). \end{cases} \tag{38}$$

By construction, $\bar{\pi}$ is a valid policy supported only on the retained actions, and hence $\bar{\pi} \in \widehat{\Pi}$.

For any policy μ, define its Bellman operator as

$$(\mathcal{T}^{\mu} V)(s) = \sum_{a \in \mathcal{A}} \mu(a \mid s) \mathbb{E}\left[R(s,a) + \gamma V(S') \mid s, a\right], \tag{39}$$

where $S' \sim P(\cdot \mid s, a)$. Because

$$Q^{\pi}(s,a) = \mathbb{E}\left[R(s,a) + \gamma V^{\pi}(S') \mid s, a\right], \tag{40}$$

the definition of $\bar{\pi}$ gives

$$(\mathcal{T}^{\bar{\pi}} V^{\pi})(s) - (\mathcal{T}^{\pi} V^{\pi})(s) = \sum_{a \in \mathcal{R}(s)} \pi(a \mid s) \left[Q^{\pi}(s, \varphi_s(a)) - Q^{\pi}(s,a)\right]. \tag{41}$$

By the Kantorovich–Rubinstein duality (Villani et al., 2009) and Eq. (18),

$$\begin{aligned} |Q^{\pi}(s, \varphi_s(a)) - Q^{\pi}(s,a)| &= \left|\mathbb{E}\left[Z^{\pi}(s, \varphi_s(a))\right] - \mathbb{E}\left[Z^{\pi}(s,a)\right]\right| \\ &\le W_1\left(Z^{\pi}(s,a), Z^{\pi}(s, \varphi_s(a))\right) \le \epsilon. \end{aligned} \tag{42}$$

Since $V^{\pi} = \mathcal{T}^{\pi} V^{\pi}$, it follows that

$$\left|(\mathcal{T}^{\bar{\pi}} V^{\pi})(s) - V^{\pi}(s)\right| \le \sum_{a \in \mathcal{R}(s)} \pi(a \mid s) \epsilon \le \epsilon. \tag{43}$$

Therefore,

$$\left\|\mathcal{T}^{\bar{\pi}} V^{\pi} - V^{\pi}\right\|_{\infty} \le \epsilon. \tag{44}$$

Using the fixed-point identity $V^{\bar{\pi}} = \mathcal{T}^{\bar{\pi}} V^{\bar{\pi}}$ and the γ-contraction of $\mathcal{T}^{\bar{\pi}}$, we obtain

$$\begin{aligned} \left\|V^{\bar{\pi}} - V^{\pi}\right\|_{\infty} &\le \left\|\mathcal{T}^{\bar{\pi}} V^{\bar{\pi}} - \mathcal{T}^{\bar{\pi}} V^{\pi}\right\|_{\infty} + \left\|\mathcal{T}^{\bar{\pi}} V^{\pi} - V^{\pi}\right\|_{\infty} \\ &\le \gamma \left\|V^{\bar{\pi}} - V^{\pi}\right\|_{\infty} + \epsilon. \end{aligned} \tag{45}$$

Rearranging gives

$$\left\|V^{\bar{\pi}} - V^{\pi}\right\|_{\infty} \le \frac{\epsilon}{1-\gamma}, \tag{46}$$

which proves Theorem 1. □

F.2 Optimal-control Specialization

Taking $\pi = \pi^{\star}$ in Theorem 1 gives

$$\left\|V^{\bar{\pi}^{\star}} - V^{\star}\right\|_{\infty} \le \frac{\epsilon}{1-\gamma}. \tag{47}$$

Because the projected policy $\bar{\pi}^{\star}$ is feasible in the masked MDP and masking only restricts the feasible policy class,

$$V^{\bar{\pi}^{\star}}(s) \le \widehat{V}^{\star}(s) \le V^{\star}(s), \qquad \forall s \in \mathcal{S}. \tag{48}$$

Combining Eqs. (47) and (48) yields

$$0 \le V^\star(s) - \widehat{V}^\star(s) \le V^\star(s) - V^{\bar{\pi}^\star}(s) \\ \le \left\| V^\star - V^{\bar{\pi}^\star} \right\|_\infty \le \frac{\epsilon}{1-\gamma}, \qquad \forall s \in \mathcal{S}, \tag{49}$$

which proves Eq. (20).

For completeness, the same result can also be established directly through the Bellman optimality operators.

Direct Bellman proof. Let $\mathcal{M}$ and $\widehat{\mathcal{M}}$ denote the original and masked MDPs, respectively, and define the retained action set at state s as

$$\widehat{\mathcal{A}}(s) = \mathcal{A} \setminus \mathcal{R}(s). \tag{50}$$

Let $\mathcal{T}$ and $\widehat{\mathcal{T}}$ be their Bellman optimality operators:

$$(\mathcal{T}V)(s) = \max_{a \in \mathcal{A}} \mathbb{E}\left[R(s,a) + \gamma V(S') \mid s, a\right], \tag{51}$$

$$(\widehat{\mathcal{T}}V)(s) = \max_{a \in \widehat{\mathcal{A}}(s)} \mathbb{E}\left[R(s,a) + \gamma V(S') \mid s, a\right]. \tag{52}$$

For each state s, choose an optimal action $a_s^\star \in \arg\max_{a \in \mathcal{A}} Q^\star(s,a)$. If $a_s^\star \notin \mathcal{R}(s)$, then it remains feasible in the masked MDP and

$$(\widehat{\mathcal{T}}V^\star)(s) = (\mathcal{T}V^\star)(s) = V^\star(s). \tag{53}$$

If $a_s^\star \in \mathcal{R}(s)$, its representative $\varphi_s(a_s^\star)$ remains feasible. By the Kantorovich–Rubinstein duality,

$$Q^\star(s, \varphi_s(a_s^\star)) \ge Q^\star(s, a_s^\star) - W_1\left(Z^\star(s, a_s^\star), Z^\star(s, \varphi_s(a_s^\star))\right) \\ \ge V^\star(s) - \epsilon. \tag{54}$$

Consequently,

$$(\widehat{\mathcal{T}}V^\star)(s) \ge Q^\star(s, \varphi_s(a_s^\star)) \ge V^\star(s) - \epsilon. \tag{55}$$

Since masking restricts the maximization, $(\widehat{\mathcal{T}}V^\star)(s) \le (\mathcal{T}V^\star)(s) = V^\star(s)$. Thus, for every state,

$$0 \le (\mathcal{T}V^\star)(s) - (\widehat{\mathcal{T}}V^\star)(s) \le \epsilon, \tag{56}$$

and hence

$$\left\| \mathcal{T}V^\star - \widehat{\mathcal{T}}V^\star \right\|_\infty \le \epsilon. \tag{57}$$

Because $V^\star = \mathcal{T}V^\star$ and $\widehat{V}^\star = \widehat{\mathcal{T}}\widehat{V}^\star$, the triangle inequality and the γ-contraction of $\widehat{\mathcal{T}}$ give

$$\left\| V^\star - \widehat{V}^\star \right\|_\infty \le \left\| \mathcal{T}V^\star - \widehat{\mathcal{T}}V^\star \right\|_\infty + \left\| \widehat{\mathcal{T}}V^\star - \widehat{\mathcal{T}}\widehat{V}^\star \right\|_\infty \\ \le \epsilon + \gamma \left\| V^\star - \widehat{V}^\star \right\|_\infty. \tag{58}$$

Rearranging yields

$$\left\| V^\star - \widehat{V}^\star \right\|_\infty \le \frac{\epsilon}{1-\gamma}. \tag{59}$$

Finally, masking restricts the feasible policy class, so $V^\star(s) \ge \widehat{V}^\star(s)$ for every state. Therefore,

$$0 \le V^\star(s) - \widehat{V}^\star(s) \le \frac{\epsilon}{1-\gamma}, \qquad \forall s \in \mathcal{S}. \tag{60}$$

□

G Proof of Corollary 1: Safe Masking under Estimation Error

G.1 Action-dependent Estimation-error Bound

Proof. Fix any action pair $(a, \varphi_s(a))$ accepted for masking. By the triangle inequality for the 1-Wasserstein distance,

$$\begin{aligned} W_1\left(Z^\pi(s,a), Z^\pi(s,\varphi_s(a))\right) \le\ & W_1\left(Z^\pi(s,a), Z_\theta^\pi(s,a)\right) \\ & + W_1\left(Z_\theta^\pi(s,a), Z_\theta^\pi(s,\varphi_s(a))\right) \\ & + W_1\left(Z_\theta^\pi(s,\varphi_s(a)), Z^\pi(s,\varphi_s(a))\right). \end{aligned} \tag{61}$$

Applying the assumptions of Corollary 1 gives

$$\begin{aligned} W_1\left(Z^\pi(s,a), Z^\pi(s,\varphi_s(a))\right) &\le \eta(s,a) + \epsilon + \eta(s,\varphi_s(a)) \\ &\le \epsilon_{\text{eff}}. \end{aligned} \tag{62}$$

Thus, every masked action and its retained representative satisfy the assumption of Theorem 1 with ϵ_{eff} in place of ϵ. Applying that theorem yields

$$\left\|V^{\bar{\pi}} - V^\pi\right\|_\infty \le \frac{\epsilon_{\text{eff}}}{1-\gamma}. \tag{63}$$

When $\pi = \pi^\star$, the corresponding optimal-control bound follows from the optimal-policy specialization in Appendix F.2. □

G.2 Uniform-error Special Case

Suppose that all local estimation errors associated with the action pairs accepted for masking satisfy a common upper bound:

$$W_1\left(Z^\pi(s,b), Z_\theta^\pi(s,b)\right) \le \eta(s,b) \le \eta, \quad \forall s \in \mathcal{S}, a \in \mathcal{R}(s), b \in \{a, \varphi_s(a)\}. \tag{64}$$

Then

$$\epsilon_{\text{eff}} \le \epsilon + 2\eta, \tag{65}$$

and Corollary 1 reduces to

$$\left\|V^{\bar{\pi}} - V^\pi\right\|_\infty \le \frac{\epsilon + 2\eta}{1-\gamma}. \tag{66}$$

The previous high-probability uniform-error result holds under the stronger global event

$$\mathcal{E}_\eta = \left\{ W_1\left(Z_\theta^{\pi^\star}(s,a), Z^\star(s,a)\right) \le \eta, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \right\}. \tag{67}$$

On the event $\mathcal{E}_\eta$, Corollary 1 applies with $\eta(s,b) = \eta$ for every action involved in a masking decision. Therefore, if $\Pr(\mathcal{E}_\eta) \ge 1-\delta$, the resulting value bound holds with probability at least $1-\delta$. Taking $\pi = \pi^\star$ in Eq. (66) and applying the optimal-control specialization gives

$$0 \le V^\star(s) - \widehat{V}^\star(s) \le \frac{\epsilon + 2\eta}{1-\gamma}, \qquad \forall s \in \mathcal{S}. \tag{68}$$

H Algorithm of DAET-Mask

Algorithm 1 presents the QR-DQN instantiation of DAET-Mask used in the GridWorld experiment. The PPO instantiation follows the same dual-gating, ROPE testing, and masking procedure, with implementation details provided in Appendix J.

I Experimental Details for Simulation Study in GridWorld

I.1 Dataset Details

We first evaluate DAET-Mask in a controlled GridWorld environment. As shown in Figure 3, we construct a 10×10 GridWorld with a fork-join topology. The agent starts from $(0,0)$ and aims to

Algorithm 1 DAET-Mask: Distributional Action Equivalence Testing with Dynamic Masking

Require: Environment $\mathcal{E}$, replay buffer $\mathcal{D}$, and a distributional RL network (e.g., QR-DQN) with K ensemble heads parameterized by $\{\theta_k\}_{k=1}^{\mathrm{K}}$.

Require: Hyperparameters: ensemble size K, number of quantiles N, network update interval f_{update}, masking interval f_{mask}, mask validity horizon T_{mask}, epistemic threshold τ_{ep}, Q-value margin Δ_Q, Wasserstein tolerance ϵ, and tolerance error level α.

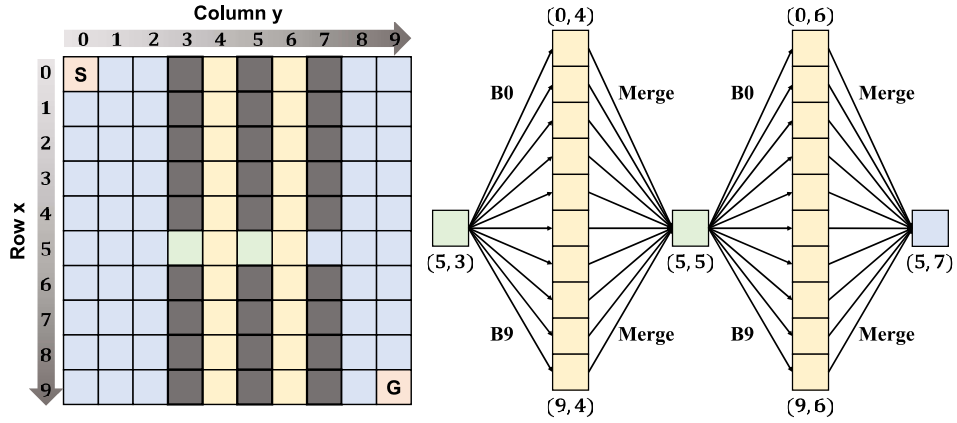
Initialize network parameters $\{\theta_k\}_{k=1}^{\mathrm{K}}$ and target network parameters $\{\theta_k^-\}_{k=1}^{\mathrm{K}} \leftarrow \{\theta_k\}_{k=1}^{\mathrm{K}}$.
Initialize the expiration dictionary $t_{\text{exp}}(s,a) \leftarrow 0$ for encountered state-action pairs if $T_{\text{mask}} > 1$.
for each training step $t = 1, 2, \ldots, T$ **do**
  // Actor: Online Interaction
  Observe current state s_t from $\mathcal{E}$.
  // Construct the current mask $\mathrm{M}_t(s_t)$
  if $T_{\text{mask}} > 1$ **then**
    Construct the current mask $\mathrm{M}_t(s_t)$ from $t_{\text{exp}}(s_t, \cdot)$:

$$\mathrm{M}_t(s_t,a) \leftarrow \begin{cases} -\infty, & t < t_{\text{exp}}(s_t,a), \\ 0, & t \geq t_{\text{exp}}(s_t,a), \end{cases} \qquad \forall a \in \mathcal{A}. \tag{69}$$

  else
    Initialize a fresh current-step mask:

$$\mathrm{M}_t(s_t,a) \leftarrow 0, \forall a \in \mathcal{A} \tag{70}$$

  end if
  Compute masked action scores using the current policy/value network.
  Select action a_t using the masked policy (e.g., ϵ-greedy or Softmax).
  Execute a_t, observe reward r_t and next state s_{t+1}, store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$.
  // Learner: Network Update
  if $t \bmod f_{\text{update}} = 0$ **then**
    Sample a mini-batch of transitions $\mathcal{B}_{\text{update}} \sim \mathcal{D}$.
    Update $\{\theta_k\}_{k=1}^{\mathrm{K}}$ using the distributional Bellman loss (e.g., Quantile Huber loss).
  end if
  // Periodic DAET and Mask Activation
  if $t \bmod f_{\text{mask}} = 0$ **then**
    Sample a mini-batch of transitions $\mathcal{B}_{\text{mask}} \sim \mathcal{D}$.
    for each distinct state $s \in \mathcal{B}_{\text{mask}}$ **do**
      Construct $\mathrm{M}_t(s)$ from $t_{\text{exp}}(s, \cdot)$.
      Define the currently available action set $\mathcal{A}_t(s) = \{a \in \mathcal{A} : t \geq t_{\text{exp}}(s,a)\}$.
      if $|\mathcal{A}_t(s)| \geq 2$ **then**
        For all $a \in \mathcal{A}_t(s)$, compute $\bar{Q}(s,a) = \frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \mathbb{E}[Z_{\theta_k}(s,a)]$.
        For all $a \in \mathcal{A}_t(s)$, compute epistemic uncertainties $\mathrm{U}_{\text{ep}}(s,a)$ using Eq. (12).
        Identify the current greedy action: $a^* = \arg\max_{a \in \mathcal{A}_t(s)} \bar{Q}(s,a)$.
        for each candidate action $a_i \in \mathcal{A}_t(s) \setminus \{a^*\}$ **do**
          // Dual-Gating Mechanism
          if $\max(\mathrm{U}_{\text{ep}}(s,a^*), \mathrm{U}_{\text{ep}}(s,a_i)) \leq \tau_{\text{ep}}$ **and** $|\bar{Q}(s,a^*) - \bar{Q}(s,a_i)| \leq \Delta_Q$ **then**
            // Uncertainty-Guided ROPE Decision Rule
            for $k = 1, \ldots, \mathrm{K}$ **do**
              Compute $W_1^{(k)}(s; a^*, a_i)$.
            end for
            Compute $\widehat{W}_{1,\text{upper}}^{(1-\alpha)}(s; a^*, a_i)$ using Eq. (16).
            if $\widehat{W}_{1,\text{upper}}^{(1-\alpha)}(s; a^*, a_i) \leq \epsilon$ **then**
              // Accept $\mathcal{H}_0$: Actions are ϵ-equivalent
              Activate the temporary mask: $\mathrm{M}_t(s,a_i) \leftarrow -\infty, t_{\text{exp}}(s,a_i) \leftarrow t + T_{\text{mask}}$.
            end if
          end if
        end for
      end if
    end for
  end if
end for



(a) Overall Topology (b) Fork-Join Mechanism

Figure 3: **Overview of the** 10×10 **fork-join Gridworld environment.** (a) shows the overall topography with normal regions in pale blue and Gadget regions in pale yellow. Obstacle Walls (gray) at $y \in \{3, 5, 7\}$ are impassable except at $x = 5$. The start state "S" is at $(0, 0)$, and the goal "G" is at $(9, 9)$. (b) A detailed visualization of the heterogenous 10-way Branching mechanism at $(5, 3)$ or $(5, 5)$. All ten distributional-equivalent actions (B0, ..., B9) lead to distinct branch cells $(x, 4)$ or $(x, 6)$ at a step cost of $r = -1$, followed by a mandatory merge action to $(5, 5)$ or $(5, 7)$.

Table 2: Immediate rewards for 17 discrete actions.

Index	Action	Reward	Transition	Role
0	Up	-1	$(x, y) \to (x-1, y), 0 < x \leq 9, y \in \{0, 1, 2, 8, 9\}$	Standard
1	Down	-1	$(x, y) \to (x+1, y), 0 \leq x < 9, y \in \{0, 1, 2, 8, 9\}$	Standard
2	Left	-1	$(x, y) \to (x, y-1), 0 \leq x \leq 9, y \in \{1, 2, 7, 8, 9\}$	Standard
3	Right	-1	$(x, y) \to (x, y+1), 0 \leq x \leq 9, y \in \{0, 1, 2, 7, 8\}$	Standard
4	R_safe	-1	$(x, y) \to (x, y+1), 0 \leq x \leq 9, y \in \{0, 1, 2, 8\}$	Duplicate
5	R_risk	-11(50%), +9(50%)	$(x, y) \to (x, y+1), 0 \leq x \leq 9, y \in \{0, 1, 2, 8\}$	Risky
6	B0	-1	$(5, 3) \to (0, 4), (5, 5) \to (0, 6)$	Branch
7	B1	-1	$(5, 3) \to (1, 4), (5, 5) \to (1, 6)$	Branch
8	B2	-1	$(5, 3) \to (2, 4), (5, 5) \to (2, 6)$	Branch
9	B3	-1	$(5, 3) \to (3, 4), (5, 5) \to (3, 6)$	Branch
10	B4	-1	$(5, 3) \to (4, 4), (5, 5) \to (4, 6)$	Branch
11	B5	-1	$(5, 3) \to (5, 4), (5, 5) \to (5, 6)$	Branch
12	B6	-1	$(5, 3) \to (6, 4), (5, 5) \to (6, 6)$	Branch
13	B7	-1	$(5, 3) \to (7, 4), (5, 5) \to (7, 6)$	Branch
14	B8	-1	$(5, 3) \to (8, 4), (5, 5) \to (8, 6)$	Branch
15	B9	-1	$(5, 3) \to (9, 4), (5, 5) \to (9, 6)$	Branch
16	Merge	-1	$(x, y) \to (5, y+1), 0 \leq x \leq 9, y \in \{4, 6\}$	Merge

reach the goal at $(9, 9)$. Three obstacle walls at rows $y \in 3, 5, 7$ (passable only at $x = 5$) partition the grid into regions connected by two cascaded fork-join gadgets. As shown in Table 2, the action space comprises 17 actions: four standard directional moves, a deterministic duplicate of right (R_safe), a risk-variant (R_risk) that induces the same transition and expected reward as right but with stochastic immediate reward ($+9$ or -11 with equal probability), ten branch actions B0, ..., B9 available exclusively at fork nodes, and a Merge action at branch states. Each non-terminal transition receives a mean reward of -1, and reaching the goal yields a terminal reward of 100. Episodes are truncated after 200 steps. This design simultaneously tests three equivalence categories: (i) exact equivalence (R $\equiv$ R_safe, identical transitions and rewards), (ii) 10-way distributional equivalence (Bi $\equiv$ Bj, $i \neq j$): each branch action leads to a distinct intermediate state, but all branches deterministically merge one step later, which is a blind spot for transition-based methods as all of these

actions yield identical return distributions, and (iii) risk-distinctness (R≠R_risk): identical expected returns but different return distributions, which is a blind spot for the Mean-Q Masking method.

Our evaluation comprises two axes: learning quality and masking correctness. For learning quality, Episodic Return measures the cumulative reward per episode, while Action Space Size reports the average number of available actions per decision step, reflecting the effective reduction achieved by masking. To assess masking correctness, we compare each agent's learned action mask against the ground-truth equivalence classes of the environment. Evaluation is performed per equivalence class: within each class, all members but one are redundant and should be masked, while one representative must be retained. A true positive (TP) is a correctly masked redundant member, a false positive (FP) occurs when masking eliminates the last representative of a class, and a false negative (FN) is a redundant member that remains unmasked. Precision, Recall, and F1-Score are derived from these counts in the standard manner.

Let $\mathcal{S}_{\text{vis}}$ denote the set of states visited by the agent during training (i.e., states with entries in the learned mask dictionary). For each state s, the environment provides a set of legal actions $\mathcal{A}_{\text{legal}}(s)$ and a ground-truth partition into equivalence classes $\mathcal{C}(s) = C_1, C_2, \ldots$, where actions within the same class induce identical return distributions. The agent's learned mask partitions actions into an active set $\mathcal{A}_{\text{act}}(s)$ (mask value $= 0$) and a masked set $\mathcal{A}_{\text{mask}}(s)$ (mask value $= -\infty$).

- **Episodic Return**. The cumulative undiscounted reward collected during episode e.
- **Action Space Size**. At each decision step t, the number of actions that are both environmentally legal and not masked is

$$n_t = \left|\mathcal{A}_{\text{legal}}(s_t) \cap \mathcal{A}_{\text{act}}(s_t)\right|. \tag{71}$$

  The episode-level metric is the per-step average $\bar{n}_e = T_e^{-1} \sum_t n_t$. For unmasked baselines (e.g., QR-DQN), $\mathcal{A}_{\text{act}}(s) = \mathcal{A}$ for all s, n_t reduces to the number of legal actions.
- **TP, FP, and FN**. Masking correctness is evaluated per equivalence class. For each visited state $s \in \mathcal{S}_{\text{vis}}$ and each class $C \in \mathcal{C}(s)$, let

$$P_C = |C \cap \mathcal{A}_{\text{masked}}(s)|, \qquad A_C = |C \cap \mathcal{A}_{\text{act}}(s)|, \qquad K = |C|. \tag{72}$$

  The per-class contributions to the global counts are defined as:

$$(\text{TP}_C, \text{FP}_C) = \begin{cases} (P_C, 0) & \text{if } K \geq 2 \text{ and } A_C \geq 1, \\ (K-1, 1) & \text{if } K \geq 2 \text{ and } A_C = 0, \\ (0, P_C) & \text{if } K = 1. \end{cases} \tag{73}$$

  The three cases correspond to the following situations: If $K \geq 2$, $A_C \geq 1$, it means at least one representative survives, thus all P_C masked members are true positives (correct masking). If $K \geq 2$, $A_C = 0$, it means the entire class has been eliminated (over masking). Of the K members pruned, $K-1$ correspond to genuinely redundant actions (TP) and one corresponds to the destroyed representative (FP). If $K = 1$, it means a non-redundant action was masked, and every such masking is a false positive (singleton violation). Global counts are obtained by summation over all visited states and classes:

$$\text{TP} = \sum_{s \in \mathcal{S}_{\text{vis}}} \sum_{C \in \mathcal{C}(s)} \text{TP}_C, \qquad \text{FP} = \sum_{s \in \mathcal{S}_{\text{vis}}} \sum_{C \in \mathcal{C}(s)} \text{FP}_C. \tag{74}$$

  False negatives are derived from the invariant. Each multi-member class contributes $K-1$ condition positives. The total number of condition positives is

$$N_+ = \sum_{s \in \mathcal{S}_{\text{vis}}} \sum_{C \in \mathcal{C}(s)} (|C| - 1), \tag{75}$$

  and the false negative count is

$$\text{FN} = N_+ - \text{TP}. \tag{76}$$

  N_+ increases only when the agent visits a previously unseen state, which accounts for the transient rise in FN during early training. As the mask accumulates, TP is monotonically non-decreasing once $\mathcal{S}_{\text{vis}}$ stabilizes. Consequently, $\text{FN} = N_+ - \text{TP}$ is monotonically non-increasing.

- **Precision, Recall, and F1-Score**. Precision measures the safety of masking decisions: among all masked actions, what fraction were genuinely redundant? Recall measures the completeness of masking: among all redundancies that should have been eliminated, what fraction were actually masked? F1-Score is their harmonic mean, balancing safety and completeness.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{N_+}, \quad \text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{77}$$

I.2 Baseline Details

We evaluate DAET-Mask against four baselines. The first is the standard QR-DQN, serving as the unmasked distributional RL reference (Dabney et al., 2018). The second is Mean-Q Masking, which masks actions solely by expected-return proximity (Definition 5) (Li et al., 2006; Abel et al., 2016). The remaining two baselines probe the transition-based equivalence criterion (Definition 4) through two learning-based variants: a Forward-Dynamics baseline that masks actions yielding identical predicted forward transitions (Abel et al., 2020; Gelada et al., 2019), and an Inverse-Dynamics baseline that identifies redundancy utilizing an inverse dynamics posterior (Baram et al., 2021; Zhong et al., 2024).

Definition 4 (Transition-based Equivalence)**.** *Two actions $a_1, a_2 \in \mathcal{A}$ are said to be transition-equivalent in state s if they induce identical transition probabilities and immediate rewards:*

$$P(\cdot|s, a_1) = P(\cdot|s, a_2) \quad \textit{and} \quad R(s, a_1) = R(s, a_2). \tag{78}$$

Definition 5 (Mean-based Equivalence)**.** *Two actions $a_1, a_2 \in \mathcal{A}$ are said to be mean-equivalent in state s under policy π if their expected returns are identical:*

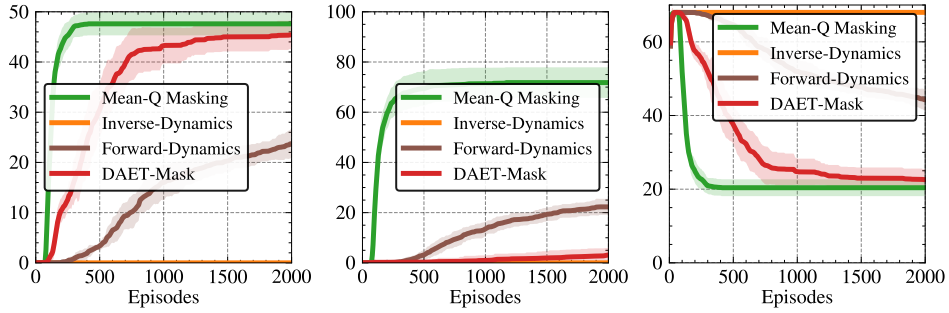
$$Q^\pi(s, a_1) = Q^\pi(s, a_2). \tag{79}$$

I.3 Experimental Settings

In this section, we detail the experimental setup for the GridWorld simulation. All experiments were conducted using Python 3.9.23 and PyTorch 2.4.1.

We employ QR-DQN with Double DQN-style target computation as the base reinforcement learning algorithm, with an ϵ-greedy exploration schedule decaying linearly from 1.0 to 0.01 over the first 1000 episodes. The agent is trained for 2000 episodes across 10 independent seeds. A replay buffer of size 100,000 is used, with a warm-up period of 1000 transitions before learning begins. Minibatches of size 64 are sampled every 4 environment steps. The discount factor is $\gamma = 0.95$, the learning rate is 5×10^{-4} (Adam optimizer), and the target network is updated via Polyak averaging with coefficient $\tau_{\text{EMA}} = 0.005$. The quantile Huber loss threshold is κ = 1.0. Gradients are clipped to a maximum norm of 10.0.

The observation is a 100-dimensional one-hot encoding of the agent's grid position. Each QR-DQN head is a 2-hidden-layer MLP ($100 \to 128 \to 128 \to \text{N} \times |\mathcal{A}|$) with ReLU activations and Kaiming normal initialization. For the DAET-Mask agent, $\text{K} = 10$ independently initialized heads form a deep ensemble (with dropout disabled), sharing a single replay buffer. Each head outputs $\text{N} = 32$ quantile atoms per action. The ensemble-mean Q-value is used for greedy action selection, while inter-head disagreement provides epistemic uncertainty estimates. The DAET-Mask pipeline uses the following hyperparameters: epistemic CV threshold $\tau_{\text{ep}} = 0.3$, Q-value margin $\Delta_Q = 0.3$, Wasserstein ROPE radius $\epsilon = 0.3$, and tolerance error level $\alpha = 0.05$. Masking begins after 1000 environment steps and is performed every 50 gradient updates. All masking baselines share the same masking schedule. Regarding the dynamic masking horizon, we set $T_{\text{mask}} = \infty$. In this controlled GridWorld environment, the state space is small and discrete ($10 \times 10 = 100$ states), meaning states are frequently revisited. Furthermore, the manually injected redundant actions are exact duplicates, rendering their distributional equivalence stationary under any continuation policy. Therefore, once an action is confidently identified as redundant at a specific state, maintaining a persistent state-indexed mask is highly efficient. During training, we record the episodic return and the average number of active (unmasked) actions per step.

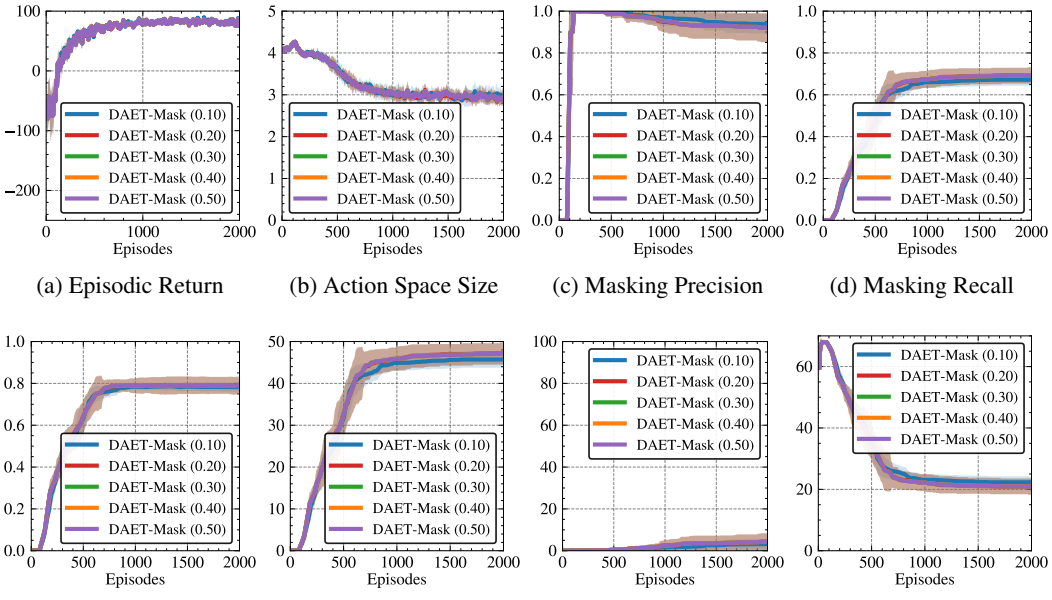


(a) TP (b) FP (c) FN

Figure 4: Supplementary raw masking metrics corresponding to Figure 2. Panels (a)–(c) detail the True Positives (TP), False Positives (FP), and False Negatives (FN) of DAET-Mask and baseline methods over training episodes in the fork-join GridWorld environment. Solid lines and shaded regions represent the mean ± std across 10 random seeds.

I.4 Decomposition of Masking Metrics: TP, FP, and FN

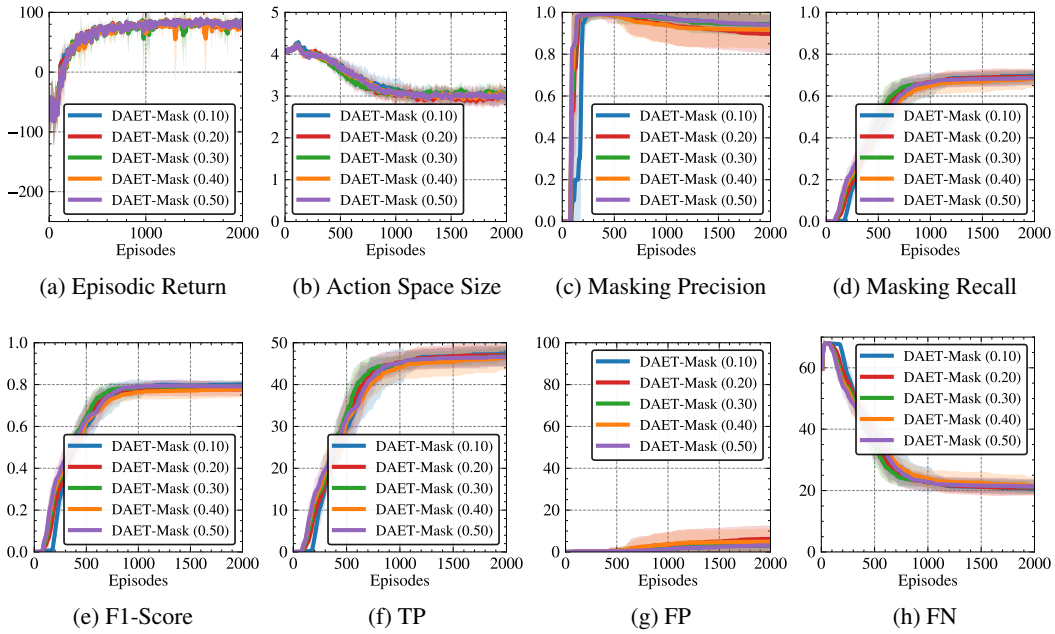
To provide deeper insights into the pruning behaviors summarized by the Precision, Recall, and F1-score in Section 4, Figure 4 presents the absolute counts of True Positives (TP), False Positives (FP), and False Negatives (FN) throughout the training process. In the context of our fork-join GridWorld environment, TP represents the number of valid (or optimal) actions correctly retained by the mask, FP denotes invalid actions erroneously kept in the actionable space, and FN indicates valid actions incorrectly masked out.



(a) Episodic Return (b) Action Space Size (c) Masking Precision (d) Masking Recall

(e) F1-Score (f) TP (g) FP (h) FN

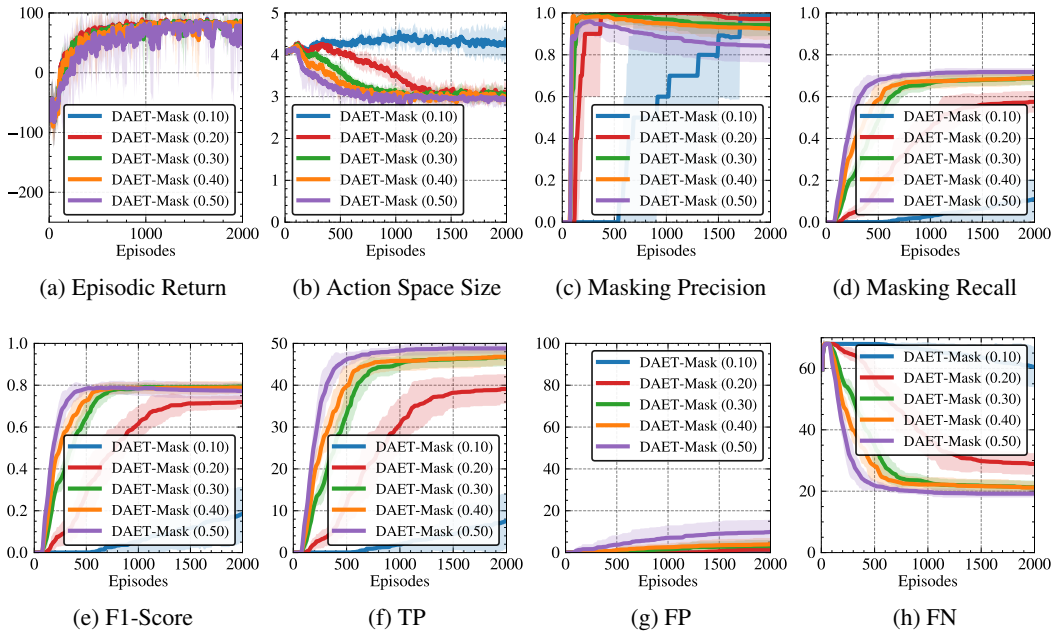
Figure 5: Sensitivity analysis of Δ_Q for DAET-Mask in GridWorld. We vary this parameter across $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ while keeping the other two key parameters fixed at the main-text setting $(\Delta_Q, \tau_{\text{ep}}, \epsilon) = (0.3, 0.3, 0.3)$. Panels (a)–(h) show different metrics over training episodes, respectively. Solid lines and shaded regions represent the mean ± std across 10 random seeds.



(a) Episodic Return (b) Action Space Size (c) Masking Precision (d) Masking Recall

(e) F1-Score (f) TP (g) FP (h) FN

Figure 6: Sensitivity analysis of τ_{ep} for DAET-Mask in GridWorld. We vary this parameter across $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ while keeping the other two key parameters fixed at the main-text setting $(\Delta_Q, \tau_{\text{ep}}, \epsilon) = (0.3, 0.3, 0.3)$. Panels (a)–(h) show different metrics over training episodes, respectively. Solid lines and shaded regions represent the mean $\pm$ std across 10 random seeds.



(a) Episodic Return (b) Action Space Size (c) Masking Precision (d) Masking Recall

(e) F1-Score (f) TP (g) FP (h) FN

Figure 7: Sensitivity analysis of ϵ for DAET-Mask in GridWorld. We vary this parameter across $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ while keeping the other two key parameters fixed at the main-text setting $(\Delta_Q, \tau_{\text{ep}}, \epsilon) = (0.3, 0.3, 0.3)$. Panels (a)–(h) show different metrics over training episodes, respectively. Solid lines and shaded regions represent the mean $\pm$ std across 10 random seeds.

I.5 Sensitivity Analysis

We further analyze the sensitivity of DAET-Mask to its three key hyperparameters: the expectation gate threshold Δ_Q, the epistemic confidence threshold τ_{ep}, and the ROPE threshold ϵ. The main-text results use $(\Delta_Q, \tau_{\text{ep}}, \epsilon) = (0.3, 0.3, 0.3)$. In each sensitivity experiment, we vary one hyperparameter while keeping the other two fixed at their default values. Specifically, we evaluate $\Delta_Q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, $\tau_{\text{ep}} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

Effect of the expectation gate threshold Δ_Q. As shown in Figure 5, DAET-Mask maintains stable learning performance across different values of Δ_Q and converges to similar episodic returns. The number of active actions also remains within a comparable range (between 2.93 and 2.98), without exhibiting monotonic over-masking as Δ_Q increases. This indicates that DAET-Mask is not overly sensitive to the expectation gate. This behavior is expected because Δ_Q only serves as a necessary-condition filter: a larger Δ_Q allows more candidate pairs to enter the subsequent distributional test, but the final masking decision is still governed by the epistemic confidence gate and the uncertainty-guided ROPE rule. Thus, even when the expectation gate is relatively permissive, DAET-Mask does not collapse into aggressive mean-based masking.

Effect of the epistemic confidence threshold τ_{ep}. As shown in Figure 6, DAET-Mask is also robust to the epistemic confidence threshold in this diagnostic GridWorld setting. The episodic-return curves are nearly indistinguishable across the tested values of τ_{ep}, and the corresponding active-action curves remain highly similar. This suggests that, once sufficient exploration has been accumulated in this relatively small environment, the epistemic confidence gate does not become the dominant factor controlling masking. Instead, most candidate pairs that are eligible for testing exhibit similar confidence behavior across the tested threshold range. Importantly, varying τ_{ep} does not destabilize learning, confirming that the confidence gate can regulate testing reliability without introducing brittle threshold sensitivity.

Effect of the ROPE threshold ϵ. As shown in Figure 7, ϵ has the most visible effect on action-space reduction. A smaller ϵ defines a narrower region of practical equivalence and therefore makes the ROPE decision more conservative. In particular, $\epsilon = 0.1$ retains almost the full action space for much of training, indicating only minimal masking. As ϵ increases, the equivalence region becomes more permissive, allowing DAET-Mask to identify more actions as practically redundant and thereby reducing the number of active actions more substantially. Despite this stronger masking, the episodic-return curves remain stable across the tested values, suggesting that the additional masking induced by larger ϵ does not degrade learning performance in this environment. The default setting $\epsilon = 0.3$ therefore provides a balanced configuration: it yields meaningful action-space reduction while preserving stable return performance.

Overall, the sensitivity analysis shows that DAET-Mask is robust across a broad range of hyperparameter values. The expectation gate Δ_Q primarily controls candidate screening, the epistemic threshold τ_{ep} regulates whether estimates are sufficiently reliable for testing, and the ROPE threshold ϵ directly controls the tolerance for practical equivalence. Across all three dimensions, DAET-Mask preserves stable return performance while maintaining effective action-space reduction, supporting the robustness of the proposed testing-based masking mechanism.

I.6 Per-Seed Results

To make the variability across independent runs transparent, Table 3 reports the final evaluation results for each of the ten random seeds $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The summary rows report the sample mean and sample standard deviation across seeds.

J Experimental Details for Real-world Study in RMFS

J.1 Dataset Details

We evaluate the real-world applicability of our method on an industrial robotic mobile fulfillment system (RMFS) dataset. The environment corresponds to a goods-to-person warehouse, where mobile robots transport inventory shelves between storage locations and workstations to fulfill dy-

Table 3: Per-seed experiment results in GridWorld.

Method	Seed	Episodic Return	Action Space Size	Precision	Recall
QR-DQN	0	82	5.28	–	–
	1	82	4.78	–	–
	2	92	4.28	–	–
	3	62	4.56	–	–
	4	72	4.44	–	–
	5	72	5.17	–	–
	6	80	4.15	–	–
	7	82	4.78	–	–
	8	92	4.72	–	–
	9	62	4.61	–	–
	Mean ± Std.	77.80 ± 10.14	4.68 ± 0.34	–	–
Mean-Q Masking	0	-200	1.00	0.45	0.72
	1	-70	1.00	0.39	0.72
	2	-200	1.50	0.39	0.69
	3	-200	1.49	0.40	0.72
	4	-210	1.99	0.38	0.62
	5	-200	1.01	0.42	0.69
	6	-200	1.00	0.40	0.74
	7	-200	1.50	0.41	0.69
	8	-190	1.00	0.40	0.74
	9	-220	1.02	0.36	0.68
	Mean ± Std.	-189 ± 40.36	1.25 ± 0.33	0.40 ± 0.02	0.70 ± 0.03
Inverse-Dynamics	0	62	4.61	0.00	0.00
	1	82	4.67	0.00	0.00
	2	92	4.50	0.00	0.00
	3	62	4.61	0.00	0.00
	4	72	4.83	0.00	0.00
	5	82	4.72	0.00	0.00
	6	82	4.06	0.00	0.00
	7	82	5.33	0.00	0.00
	8	92	4.06	0.00	0.00
	9	52	4.78	0.00	0.00
	Mean ± Std.	76 ± 12.81	4.62 ± 0.35	0.00 ± 0.00	0.00 ± 0.00
Forward-Dynamics	0	82	3.50	0.47	0.32
	1	92	3.67	0.52	0.41
	2	102	3.83	0.57	0.41
	3	82	3.89	0.48	0.31
	4	82	3.78	0.50	0.40
	5	82	3.78	0.52	0.34
	6	82	3.89	0.43	0.28
	7	82	3.94	0.55	0.32
	8	92	3.83	0.56	0.34
	9	78	3.77	0.56	0.35
	Mean ± Std.	85.60 ± 6.97	3.79 ± 0.12	0.52 ± 0.04	0.35 ± 0.04
DAET-Mask	0	82	2.70	0.90	0.68
	1	97	2.94	1.00	0.65
	2	82	3.33	0.96	0.69
	3	77	2.89	0.89	0.72
	4	87	2.89	0.83	0.74
	5	87	3.06	0.94	0.65
	6	82	3.22	0.93	0.63
	7	82	2.78	0.98	0.65
	8	102	2.94	1.00	0.69
	9	87	3.06	0.95	0.59
	Mean ± Std.	86.50 ± 7.23	2.98 ± 0.18	0.94 ± 0.05	0.67 ± 0.04

Table 3: Per-seed experiment results in GridWorld (Continued).

Method	Seed	F1	TP	FP	FN
Mean-Q Masking	0	0.55	49	60	19
	1	0.51	49	77	19
	2	0.50	47	74	21
	3	0.52	49	73	19
	4	0.47	42	70	26
	5	0.52	47	65	21
	6	0.52	50	74	18
	7	0.51	47	68	21
	8	0.52	50	74	18
	9	0.47	46	83	22
	Mean $\pm$ Std.	0.51 ± 0.02	47.60 ± 2.29	71.80 ± 6.10	20.40 ± 2.29
Inverse-Dynamics	0	0.00	0	0	68
	1	0.00	0	0	68
	2	0.00	0	0	68
	3	0.00	0	0	68
	4	0.00	0	0	68
	5	0.00	0	0	68
	6	0.00	0	0	68
	7	0.00	0	0	68
	8	0.00	0	0	68
	9	0.00	0	0	68
	Mean $\pm$ Std.	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	68.00 ± 0.00
Forward-Dynamics	0	0.38	22	25	46
	1	0.46	28	26	40
	2	0.48	28	21	40
	3	0.38	21	23	47
	4	0.44	27	27	41
	5	0.41	23	21	45
	6	0.34	19	25	49
	7	0.41	22	18	46
	8	0.42	23	18	45
	9	0.43	24	19	44
	Mean $\pm$ Std.	0.41 ± 0.04	23.70 ± 2.90	22.30 ± 3.20	44.30 ± 2.90
DAET-Mask	0	0.77	46	5	22
	1	0.79	44	0	24
	2	0.80	47	2	21
	3	0.80	49	6	19
	4	0.78	50	10	18
	5	0.77	44	3	24
	6	0.75	43	3	25
	7	0.78	44	1	24
	8	0.82	47	0	21
	9	0.73	40	2	28
	Mean $\pm$ Std.	0.78 ± 0.02	45.40 ± 2.84	3.20 ± 2.93	22.60 ± 2.84

namically arriving customer orders. Each decision instance requires the agent to coordinate order allocation and robot scheduling under real-time operational constraints.

The dataset is constructed from real industrial warehouse data, including the physical warehouse layout, inventory snapshots, and historical order streams. The warehouse is represented as a 40×72 grid and contains 861 shelves and 16 workstations. The problem instances are configured with a fleet of 15 mobile robots tasked with fulfilling a stream of 200 dynamically arriving orders. The historical order data span 31 days and are chronologically split into training, validation, and test periods of 21, 5, and 5 days, respectively. To focus evaluation on non-trivial scheduling regimes, problem instances are sampled from periods with high order density. The validation and test sets each contain 100 instances. The time required for a worker to process items at a workstation is formulated as $\texttt{num_item} \times C_{\text{item}} + C_{\text{shelf}}$, where the per-item processing time C_{item} is set to 2, and

the base shelf processing time C_{shelf} is set to 5. This dataset therefore provides a realistic evaluation setting for assessing whether action-space reduction can improve decision efficiency in large-scale, dynamic warehouse scheduling.

J.2 Default Order Allocation Strategy

Algorithm 2 Default Order Allocation

Require: Arriving order o; Set of workstations $\mathcal{W}$; Set of shelves $\mathcal{S}$;
Ensure: The workstation allocated to order w_o; The shelves set allocated to order $\mathcal{S}_o$

```
Select allocated workstation w_o based on rules
S_o ← ∅
while d_o ≠ 0 do
    for each shelf s ∈ S do
        Calculate matching degree V^(o)_{s,w} = (Σ_{i=1}^{N_k} min(d_o[i], q_s[i])) / (dist(s, w) + ε)
    end for
    Select the shelf s' with the maximum matching degree
    S_o ← S_o ∪ {s'}
    a ← min(d_o, q_{s'})
    d_o ← d_o − a
    q_{s'} ← q_{s'} − a
end while
return w_o, S_o
```

The default order allocation strategy is described in Algorithm 2, which is employed in the following baselines.

J.3 Baseline Details

We selected two categories of methodologies: phased decision-making methods and joint decision-making methods. For the phased methods, we selected three order allocation methods and four robot scheduling methods, respectively, and evaluated the performance of all nine resulting combinations.

Order allocation methods.

- **Shortest Queue First (SQF) (Liang et al., 2023)**: This method allocates the order to the workstation with the shortest expected queue time and subsequently selects shelves based on Algorithm 2.
- **Work Load Balancing (WLB) (Tadumadze et al., 2023)**: This method allocates the order to the workstation with the minimum workload and subsequently selects shelves based Algorithm 2.
- **OR Tools**: This method model the dynamic order allocation as a static problem via a rolling-horizon approach and solve it using the Google OR-Tools CP-SAT solver.

Robot scheduling methods.

- **Nearest Neighbor (Ma et al., 2017)**: This method selects the nearest valid target.
- **Earliest Arrival Order (Deng et al., 2016)**: This method selects the earliest arriving order for completion. Furthermore, to adapt it to our three-phase decision problem, its return strategy is set to the nearest neighbor approach.
- **TSP (Li et al., 2024)**: This method models the robot scheduling process as a Traveling Salesman Problem.
- **PSMDRL (Zhou et al., 2024)**: A reinforcement learning method employs a Transformer architecture for decision.

Joint optimization methods.

- **JOTP (Zhang et al., 2025)**: A joint optimization method combining the Kuhn-Munkres algorithm (Munkres, 1957) and reinforcement learning.

- **SABS (Yang et al., 2021)**: A hybrid heuristic algorithm combining simulated annealing and beam search for the joint optimization.
- **SOAR (Tang et al., 2026)**: A deep reinforcement learning framework that achieves real-time joint optimization of order allocation and robot scheduling by integrating a soft order allocation mechanism with an Event-Driven Markov Decision Process (ED-MDP).

J.4 Experimental Settings

In this section, we detail the experimental setup. All experiments were conducted using Python 3.9.23 and PyTorch 2.4.1.

Regarding the reinforcement learning algorithm, we employ Proximal Policy Optimization (PPO) as the on-policy training algorithm, augmented with an auxiliary Quantile Regression loss to maintain distributional information across training. The total loss takes the form $\mathcal{L} = \mathcal{L}_{\text{PPO}} + c_v\mathcal{L}_V + c_q\mathcal{L}_{\text{QR}} - c_e\mathcal{H}$, where $\mathcal{L}_{\text{PPO}}$ is the clipped surrogate loss, $\mathcal{L}_V$ is the scalar value loss, $\mathcal{H}$ is the entropy bonus, and $\mathcal{L}_{\text{QR}}$ is the auxiliary Quantile Huber distributional TD loss (with threshold $\kappa = 1.0$) used to train the action-conditioned quantile heads. We adopt time-aware discounting $\gamma^{\Delta\tau}$ following the SMDP formulation in Tang et al. (2026), where $\gamma = 0.999$ and $\Delta\tau$ is the physical time elapsed between consecutive decision events. A separate target network, updated via exponential moving average ($\tau_{\text{EMA}} = 0.005$), provides stable bootstrap targets for the auxiliary quantile-regression loss. The number of local parallel environments is set to 8, with each environment collecting 128 time steps of data prior to each policy update. Upon the completion of data collection, the policy network undergoes updates for 4 epochs, with the data partitioned into 4 minibatches. The PPO clip coefficient is set to $\epsilon_{\text{clip}} = 0.1$, the entropy coefficient to $c_e = 0.01$, the value function coefficient to $c_v = 0.1$, and the QR loss coefficient to $c_q = 1.0$. The maximum gradient norm is limited to 0.5, and a target KL divergence threshold of 0.01 is used for early stopping within each epoch. The total number of timesteps is set to 4×10^6. During the training process, we monitored the model's performance on the validation set and retained the best-performing model as the final model.

Regarding the model architecture, we employ a Heterogeneous Graph Attention Network (HGT) as the shared backbone encoder, comprising 2 layers of heterogeneous attention convolutions with hidden dimension 128 and 4 attention heads. On top of this shared backbone, $\mathrm{K} = 10$ independently initialized quantile heads each produce $\mathrm{N} = 64$ quantile atoms per action candidate, yielding an ensemble return distribution of shape (K, N) for each state-action pair. The ensemble disagreement (coefficient of variation across heads) captures epistemic uncertainty, while the intra-head quantile spread captures aleatoric uncertainty. The heuristic candidate filter retains up to $K_{\text{filter}} = 50$ vacancy candidates per decision event.

Regarding the dynamic masking horizon, we set $T_{\text{mask}} = 1$ decision step. More generally, DAET-Mask supports both persistent and transient masking depending on the structure of the state space. Persistent state-indexed masks are suitable for discrete or frequently revisited states, while transient masks are preferable when exact state recurrence is rare. The RMFS state is a high-dimensional, event-dependent combination of robots, orders, shelves, inventories, and workstation queues. Its graph-structured state and event-dependent candidate sets are almost never repeated exactly. Consequently, maintaining a persistent state-indexed mask table would introduce substantial storage and lookup overhead while being frequently unused. We therefore construct a transient mask from the current state at every decision step, apply it exclusively to that step's action selection, and immediately expire it ($T_{\text{mask}} = 1$). All candidate actions are reconsidered at the next decision event. This implementation accommodates the asynchronous nature of RMFS and avoids carrying masking decisions across distinct high-dimensional states.

J.5 Sensitivity Analysis

We further conduct a sensitivity analysis of DAET-Mask on the RMFS task to examine whether its performance is robust to different hyperparameter choices. We vary the three key hyperparameters: the expectation gate threshold Δ_Q, the epistemic confidence threshold τ_{ep}, and the ROPE threshold ϵ. Specifically, we consider

$$\tau_{\text{ep}} \in \{0.010, 0.015, 0.025\}, \quad \Delta_Q \in \{0.1, 0.2, 0.5\}, \quad \epsilon \in \{0.1, 0.2, 0.5\}. \tag{80}$$

These choices yield $3 \times 3 \times 3 = 27$ possible configurations.

Table 4: Sensitivity analysis of DAET-Mask on the RMFS task. We vary $\Delta_Q \in \{0.1, 0.2, 0.5\}$, $\tau_{\text{ep}} \in \{0.010, 0.015, 0.025\}$, and $\epsilon \in \{0.1, 0.2, 0.5\}$. Only configurations satisfying $\Delta_Q \geq \epsilon$ are reported, since this condition preserves the theoretical validity of the expectation gate as a necessary-condition filter. Lower values are better for Obj and CompT.

$\tau_{\text{ep}} = 0.010$	$\Delta_Q = 0.1$			$\Delta_Q = 0.2$			$\Delta_Q = 0.5$		
Metrics	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$
Obj (↓)	394.84	–	–	396.96	396.85	–	396.27	395.20	396.09
CompT (↓)	146.02	–	–	146.63	146.93	–	148.24	144.44	146.95
$\tau_{\text{ep}} = 0.015$	$\Delta_Q = 0.1$			$\Delta_Q = 0.2$			$\Delta_Q = 0.5$		
Metrics	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$
Obj (↓)	397.24	–	–	398.58	395.73	–	395.61	395.22	397.61
CompT (↓)	147.25	–	–	147.58	146.19	–	145.82	144.99	146.01
$\tau_{\text{ep}} = 0.025$	$\Delta_Q = 0.1$			$\Delta_Q = 0.2$			$\Delta_Q = 0.5$		
Metrics	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$
Obj (↓)	395.44	–	–	395.85	395.54	–	396.99	395.31	396.00
CompT (↓)	146.64	–	–	146.14	148.64	–	145.66	144.91	144.35

However, as discussed in the dual-gating mechanism, the expectation gate is theoretically grounded only when $\Delta_Q \geq \epsilon$. This condition ensures that Gate 1 acts as a safe necessary-condition filter rather than a heuristic screening rule. If $\Delta_Q < \epsilon$, an action pair whose Wasserstein distance is within the ROPE threshold could still be discarded before the distributional test because its expectation gap exceeds Δ_Q. Such a configuration would make the expectation gate overly restrictive and would no longer preserve the theoretical guarantee derived from the Kantorovich–Rubinstein inequality. Therefore, we report only the 18 configurations satisfying $\Delta_Q \geq \epsilon$ and omit the remaining 9 configurations that fall outside the theoretically justified regime.

Table 4 reports the full sensitivity results over the valid configurations. Overall, DAET-Mask exhibits stable performance across all tested settings. The values of Obj and CompT remain within a narrow range, indicating that the method is not overly sensitive to the precise choice of Δ_Q, τ_{ep}, or ϵ in the RMFS environment. More importantly, all valid configurations outperform the strongest baseline reported in the main RMFS experiment, demonstrating that the performance improvement is not an artifact of a single carefully tuned hyperparameter setting.

The results also align with the roles of the three hyperparameters. The expectation threshold Δ_Q controls which candidate action pairs are admitted into the distributional equivalence test; the epistemic threshold τ_{ep} determines whether the learned return estimates are reliable enough for testing; and the ROPE threshold ϵ specifies the tolerance for practical distributional equivalence. Despite these different functions, the empirical results show that DAET-Mask maintains strong task performance across a broad range of theoretically valid configurations. This further supports the robustness of the proposed testing-based dynamic masking mechanism in realistic large-scale scheduling tasks.

J.6 Per-Instance Results and Statistical Significance Tests

To provide full transparency, Table 5 reports the mean order completion time (CompT) of SOAR and DAET-Mask on each of the 100 test instances. DAET-Mask attains a lower CompT on 79 out of 100 instances.

Moreover, we conduct two paired significance tests over the 100 test instances to verify that the improvement of DAET-Mask over SOAR is not attributable to chance. For each instance i, we compute the mean CompT under both methods and form the paired difference

$$\delta_i = \text{CompT}_i^{\text{SOAR}} - \text{CompT}_i^{\text{DAET-Mask}}. \tag{81}$$

A one-sided paired t-test with the alternative hypothesis that DAET-Mask achieves a lower mean CompT than SOAR yields $t(99) = 6.01$, $p = 1.53 \times 10^{-8}$, rejecting the null hypothesis of equal means. As a non-parametric alternative that does not assume normality of the paired differences, the one-sided Wilcoxon signed-rank test gives $W = 4251.00$, $p = 1.47 \times 10^{-9}$. Both tests confirm that DAET-Mask significantly outperforms SOAR at any conventional significance level ($p \ll 0.01$).

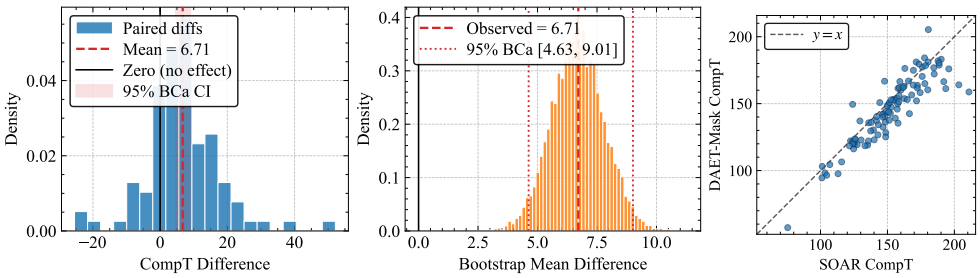
Table 5: Per-instance mean order completion time (CompT) for SOAR and DAET-Mask across 100 test instances. Lower is better. Bold entries indicate the better method for each instance.

Method	#0	#1	#2	#3	#4	#5	#6	#7	#8	#9
SOAR	147.1	135.1	103.3	166.4	125.5	**163.0**	127.3	180.1	124.7	**160.1**
DAET-Mask	**142.5**	**119.0**	**98.2**	**142.5**	**119.7**	176.3	**119.3**	**160.3**	**116.2**	160.2
Method	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19
SOAR	**180.5**	148.9	**177.7**	139.6	137.5	203.3	**159.3**	122.4	**157.9**	**147.8**
DAET-Mask	205.3	**144.1**	184.1	**130.4**	**129.6**	**164.0**	166.5	**122.2**	161.4	166.8
Method	#20	#21	#22	#23	#24	#25	#26	#27	#28	#29
SOAR	106.9	147.9	125.0	124.2	168.1	141.8	115.8	152.8	130.3	**180.4**
DAET-Mask	**104.5**	**140.5**	**120.8**	**121.7**	**156.5**	**137.3**	**106.5**	**147.3**	**122.8**	181.1
Method	#30	#31	#32	#33	#34	#35	#36	#37	#38	#39
SOAR	**162.5**	**101.3**	142.0	148.4	148.8	150.5	**158.8**	184.6	152.2	**129.4**
DAET-Mask	163.0	103.1	**140.4**	**136.3**	**128.3**	**149.3**	161.9	**176.8**	**142.7**	137.0
Method	#40	#41	#42	#43	#44	#45	#46	#47	#48	#49
SOAR	150.9	140.1	149.9	146.8	135.8	125.8	142.9	75.5	154.3	155.1
DAET-Mask	**143.8**	**123.0**	**148.2**	**132.5**	**122.1**	**123.2**	**126.7**	**57.4**	**153.2**	**152.5**
Method	#50	#51	#52	#53	#54	#55	#56	#57	#58	#59
SOAR	155.9	157.3	**144.3**	**161.7**	**169.4**	**124.0**	146.2	151.4	171.7	157.4
DAET-Mask	**139.1**	**145.9**	150.8	162.1	174.9	149.5	**141.9**	**137.6**	**152.1**	**153.9**
Method	#60	#61	#62	#63	#64	#65	#66	#67	#68	#69
SOAR	190.2	174.4	133.2	158.7	162.7	187.9	**176.1**	**171.5**	146.5	195.8
DAET-Mask	**183.2**	**158.2**	**118.5**	**155.9**	**153.7**	**182.0**	177.3	181.1	**140.0**	**176.0**
Method	#70	#71	#72	#73	#74	#75	#76	#77	#78	#79
SOAR	210.9	101.1	104.6	157.6	157.9	148.7	172.4	188.7	**165.7**	193.6
DAET-Mask	**158.7**	**94.6**	**96.5**	**149.8**	**135.4**	**125.2**	**163.6**	**178.8**	166.9	**161.2**
Method	#80	#81	#82	#83	#84	#85	#86	#87	#88	#89
SOAR	178.5	**173.6**	136.0	177.9	151.5	137.9	172.3	176.0	158.4	188.5
DAET-Mask	**178.4**	179.5	**135.7**	**165.0**	**150.7**	**136.0**	**161.8**	**170.3**	**151.4**	**180.3**
Method	#90	#91	#92	#93	#94	#95	#96	#97	#98	#99
SOAR	169.3	192.1	125.7	**152.6**	113.1	124.1	143.2	185.0	121.6	164.8
DAET-Mask	**163.9**	**166.4**	**123.7**	152.8	**97.7**	**120.0**	**134.2**	**172.1**	**118.5**	**153.1**

Figure 8a shows the distribution of the 100 paired differences δ_i. The distribution is centered well above zero (mean $\bar{\delta} = 6.71$), and the 95% BCa bootstrap confidence interval $[4.63, 9.01]$ excludes zero entirely, corroborating the parametric and non-parametric test results. Figure 8b further illustrates this by displaying the bootstrap distribution of the mean difference over 10,000 resamples: the entire distribution lies above zero, providing non-parametric confirmation that the observed improvement is robust to sampling variability. Figure 8c provides a per-instance comparison: each point represents one test instance, with the SOAR CompT on the horizontal axis and the DAET-Mask CompT on the vertical axis. The majority of points (79 out of 100) fall below the $y = x$ diagonal, confirming that DAET-Mask consistently achieves lower completion times across instances of varying difficulty.

K Discussion

DAET-Mask introduces additional computational overhead due to Deep Ensembles and pairwise distributional testing. Although the proposed dual-gating mechanism substantially reduces the number of candidate comparisons, the method is still more expensive than standard single-network RL baselines. This overhead is expected, as DAET-Mask trades additional computation for more reliable equivalence decisions and safer action-space reduction. Importantly, the empirical results sug-



(a) Paired Difference Distribution (b) Bootstrap Distribution (c) Paired Comparison

Figure 8: Paired statistical analysis between SOAR and DAET-Mask in the RMFS experiment. (a) Distribution of paired differences δ_i across 100 test instances. The dashed red line marks the mean difference $\bar{\delta} = 6.71$, and the shaded region indicates the 95% BCa bootstrap confidence interval $[4.63, 9.01]$, which excludes zero. (b) Bootstrap distribution of the mean difference obtained from 10,000 resamples. The dotted red lines delimit the 95% BCa confidence interval, and the entire distribution lies above zero. (c) Per-instance paired comparison. Each point represents one of the 100 test instances. Points below the $y = x$ diagonal indicate instances where DAET-Mask achieves lower CompT. DAET-Mask outperforms SOAR on 79 out of 100 instances.

gest that this overhead remains within an acceptable range in practice. In the RMFS experiments, DAET-Mask incurs a higher per-instance computation time than some lightweight heuristics and RL baselines, but it remains at the second-level scale and is substantially faster than optimization-based methods whose running time reaches the minute level. Moreover, this additional cost is accompanied by improved objective value, lower average completion time, and reduced effective action-space ratio. These results indicate that the computational overhead introduced by distributional testing is practically manageable in realistic scheduling scenarios, especially when decision quality and exploration efficiency are critical.

In the RMFS experiments, we adopt a shared-backbone multi-head ensemble to balance epistemic uncertainty estimation against computational efficiency. Because the ensemble heads share the same feature representation, their predictions may be correlated and less diverse than those of independently trained ensemble members. Consequently, their disagreement may underestimate epistemic uncertainty, potentially leading DAET-Mask to accept equivalence and mask actions more readily than warranted. Developing more diverse ensemble architectures or explicit uncertainty calibration methods while retaining computational efficiency is an important direction for future work.

Moreover, the current formulation of DAET-Mask is designed for discrete action spaces, where actions can be explicitly compared and masked. Extending DAET-Mask to continuous action domains is a promising direction for future work. In such settings, equivalence would need to be defined over local action neighborhoods or learned action representations, and masking may need to be replaced by density shaping, action clustering, or constrained sampling.

L Related Work

Action masking, elimination, and large action-space reduction. In many practical reinforcement learning applications, large and high-dimensional action spaces severely impede exploration efficiency and slow down policy convergence. A common practice to mitigate this is action masking, which filters out invalid actions to restrict the set of selectable actions. In large-scale game environments such as StarCraft II and Dota 2, masking is widely used to exclude actions that are invalid under game rules, cooldown constraints, target availability, or environment-provided valid-action sets (Vinyals et al., 2019; Berner et al., 2019; Huang & Ontañón, 2022). Empirical studies in real-time strategy environments further show that invalid-action masks can substantially reduce the effective action space by exploiting known game rules (Huang et al., 2021). More broadly, action-space shaping modifies the original action space by removing, merging, or discretizing actions, often relying on domain engineering to make learning tractable (Kanervisto et al., 2020). However, these masking techniques heavily rely on domain-specific and hard-coded rules.

Beyond rule-based masks, several lines of work aim to reduce large action spaces in a more algorithmic manner. Action-elimination methods remove provably or empirically suboptimal actions using confidence bounds or learned elimination signals (Even-Dar et al., 2006; Zahavy et al., 2018). Action-representation and action-embedding methods exploit structure among actions to improve generalization over large discrete action sets (Pazis & Parr, 2011; Dulac-Arnold et al., 2012; Tennenholtz & Mannor, 2019; Chandak et al., 2019). More recent masking methods attempt to learn state-specific or no-prior masks from data (Wang et al., 2024; Zhong et al., 2024). These approaches are closely related to our goal of reducing redundant exploration. However, they primarily treat action-space reduction as filtering, elimination, or representation learning directly: actions are removed because they are invalid or suboptimal. They do not generally define redundancy as an explicit state-conditioned equivalence relation among feasible actions. In contrast, DAET-Mask treats masking as the operational consequence of an uncertainty-aware equivalence test: an action is masked only when its learned return distribution is sufficiently close to that of a retained action under an uncertainty-guided ROPE criterion.

Action redundancy and equivalence in MDPs. Action redundancy concerns when distinct actions, or state-action pairs, can be treated as inducing equivalent decision consequences. This notion underlies action abstraction, which reduces the effective decision space by grouping or reusing actions according to an equivalence criterion. Classical abstraction theory is closely related to MDP model minimization, stochastic bisimulation, state abstraction, and MDP homomorphisms, which seek compact representations that preserve optimal or near-optimal behavior (Givan et al., 2003; Ravindran, 2004; Ferns et al., 2004; Li et al., 2006). From the perspective of action equivalence, existing criteria can be broadly organized into two paradigms.

The first paradigm is transition-based equivalence, which identifies actions or state-action pairs through their immediate rewards and transition effects. Bisimulation-style abstractions and MDP homomorphisms fall into this category: they provide strong theoretical guarantees, but they are restrictive because they require access to, or accurate estimation of, model-level transition structure (Givan et al., 2003; Ravindran, 2004; Ferns et al., 2004). There is prior work demonstrating that action entropy can be misleading when multiple actions induce the same or similar next-state transitions, and defining redundancy through transition entropy and action posterior estimates (Baram et al., 2021). Recent online abstraction methods similarly attempt to construct homomorphic or bisimulation-based representations from experience by matching transition distributions in latent spaces (Gelada et al., 2019; Biza & Platt, 2019; Mavor-Parker et al., 2022). These works directly motivate the need to reason about redundant actions, but their equivalence criteria remain fundamentally tied to transition effects rather than long-term return distributions.

The second paradigm is mean-based equivalence, which abstracts states or actions according to scalar value quantities, such as preserving optimal values, optimal actions, or expected returns (Li et al., 2006; Abel et al., 2016; 2020). Related action-elimination methods use confidence intervals around value or Q-function estimates to remove actions that are unlikely to be optimal (Even-Dar et al., 2006), while state-action similarity methods propagate updates across pairs expected to have similar returns (Rosenfeld et al., 2017). Such criteria are less restrictive than exact transition matching, but they collapse the full return distribution into scalar expectations or value estimates. Consequently, they cannot distinguish actions with identical means but different variability, tail behavior, or risk profiles. More importantly, both transition-based and mean-based abstractions usually treat equivalence as a property of known or accurately estimated quantities. In online RL, however, action equivalence must be inferred from finite data, where learned estimates may be unreliable in poorly explored regions. DAET addresses this gap by defining action equivalence over long-term return distributions and by accepting equivalence only through an uncertainty-aware statistical test.