CAST: Certifiable Aggregation of Smoothed Teachers for Robust Policy Adaptation

Zhongrui Zhao[1,2]* **Yanan Cai**[1,2]* **Zhigang Lu**[2] **Longkun Guo**[3]†
Ickjai Lee[1] **Shuchao Pang**[4] **Minhui Xue**[5]
[1]James Cook University, Australia [2]Western Sydney University, Australia
[3]Fuzhou University, China [4]Nanjing University of Science and Technology, China
[5]CSIRO, Australia

Abstract

Transfer learning (TL)-based policy adaptation in deep reinforcement learning (DRL) usually relies on target-side data information for retraining when facing new tasks. However, two important issues remain underexplored in practical scenarios: existing DRL transfer learning methods usually lack theoretical guarantees against adversarial attacks, and target-side data information may be unavailable for retraining. In this paper, we propose Certifiable Aggregation of Smoothed Teachers (CAST), adapting source policies by aggregating multiple certified smoothed teachers without any form of retraining, and further certifying the robustness of the adapted policy at both the action level and the cumulative-reward level. We identify three key challenges: (i) value-function transfer in DRL cannot preserve certified robustness without robust retraining; (ii) reward changes break the direct reusability of source teachers' action-level certificates; and (iii) existing certified robustness transfer results provide limited guidance for certifying the student's cumulative reward. These challenges prevent us from directly using existing TL techniques in DRL. To address them, CAST constructs robustness signals from source certificates, incorporates them into policy aggregation to obtain action-level certificates, and bridges the adapted policy's action-level certificates to its target-task cumulative reward certificate. Experiments on five multi-objective DRL benchmarks show that CAST exceeds the best source teacher in 17 out of 40 attack configurations and largely remains within the performance range of the source teachers. When testing whether smoothing benefits can be transferred, CAST improves in 39 out of 40 configurations, with a maximum relative gain of 183.9%.

1 Introduction

Deep reinforcement learning (DRL) has been widely applied to sequential decision-making tasks, and in real factory deployments, multiple DRL agents may be trained for different operational requirements [1–3]. For example, in autonomous mobile robot navigation, a factory may already have multiple certified teacher policies trained for different operational requirements: one teacher prioritizes safe navigation, another teacher prioritizes efficient delivery, and a third teacher prioritizes smooth motion [1]. When a new task requirement changes the target objective, the student policy may need to adapt to a new trade-off among them, but it may have no target-side data and cannot collect new interaction data because online exploration can interrupt production or cause safety risks. This motivates studying zero-shot transfer learning or mixture-of-experts (MoE) policy aggregation to enable policy adaptation to a new target objective without any form of retraining, instead of merely

*Equal contribution.
†Corresponding author.

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

performing policy transfer. However, although both transfer learning and MoE provide mechanisms for reusing or composing knowledge from existing policies, they are typically optimized for target-side utility rather than for preserving formal guarantees after policy transfer or composition [4–8]. As a result, even when the transferred or composed policy performs well empirically, it remains unclear whether any certified robustness property is retained, which may lead to catastrophic failures in DRL [9].

This limitation motivates us to study a certifiable transfer method for DRL without retraining. By analyzing the existing zero-shot transfer learning-based (or MoE-based) DRL and the theoretical guarantees in existing certified robustness methods, we identify *three key challenges*.

First, action- or state-value function transfer in DRL cannot preserve certified robustness without robust retraining. A key feature of DRL is that deep neural networks (DNNs) are used to approximate action-value or state-value functions for high-dimensional decision-making tasks [10]. However, studies on robustness transfer in DNNs show that certified robustness is not automatically preserved after transfer, and that robust fine-tuning or robust retraining is necessary to avoid robustness forgetting [11]. Yet directly applying such robust retraining to DRL conflicts with this motivation. Therefore, we have to exploit the DRL-specific structure of certified policies to transfer.

Second, potential reward adjustments in the student's environment break the reusability of teachers' action-level certificates. Existing certified robustness guarantees in DRL are usually tied to the policy, reward function, and certification procedure used in the original task [12–14]. When the target reward changes in the student's environment, these dependencies make it hard to directly reuse the source certificate to certify a transferred target policy. This issue is common in value-gap-based discrete certification methods [12, 13], where the certificate depends on the gap between the selected action and competing actions under the certified policy and reward. Under a changed target reward, the top action of the source policy may no longer have a positive gap, and another action may become the preferred action. Therefore, we need to certify action-level robustness based on the transferred actions.

Third, existing certified robustness transfer results provide limited guidance for certifying the student's cumulative reward. Existing cumulative reward certificates in DRL are trajectory dependent [9, 12, 15], but the student's trajectories are unknown in our target-side data-free setting. Thus, it is difficult to directly relate teacher trajectory distributions to student trajectory distributions for certification. Meanwhile, certified robustness transfer for DNNs mainly provides point-wise input output guarantees [11, 16], which correspond more closely to action-level certificates in DRL and do not certify cumulative reward over sequential interactions. Therefore, CAST requires a new framework that derives student cumulative reward certificates from the student's own action level certificates, rather than estimating action and reward certificates separately.

Building on the above, and given that randomized smoothing is a commonly used certification technique in both classification [17–19] and DRL [13], we propose *Certifiable Aggregation of Smoothed Teachers* (CAST), a robustness-aware aggregation framework that leverages randomized smoothing to certify teacher policies and guide teachers' aggregation for policy adaptation to a new task without retraining. CAST combines teacher utility with certified robustness signals, so that useful teacher decisions are transferred while teachers with stronger local certificates receive larger influence. Fig. 1 provides an overview of the CAST workflow, from certified smoothed teachers through discrete- and continuous-action aggregation to the resulting student policy. In particular, for discrete action spaces, CAST aggregates smoothed teacher action values through an advantage-margin weighted log-sum-exp rule; for continuous action spaces, CAST aggregates quantile-smoothed teacher actions through peer-review mixture weights. We further derive action-level certificates for the aggregated student and connect these certificates to a cumulative reward lower bound. Our main contributions are summarized as follows:

- We introduce CAST, a certifiable aggregation framework for training-free policy adaptation in DRL. To the best of our knowledge, CAST is the first framework that adapts multiple certified smoothed teacher policies to a new target preference without target-side data, target environment interaction, or additional retraining.
- We develop a new certification analysis for aggregated student policies. To the best of our knowledge, this is the first analysis of how state-dependent aggregation weights change under adversarial state perturbations, and the first to connect action-level certification with cumulative reward certification for aggregated student policies.

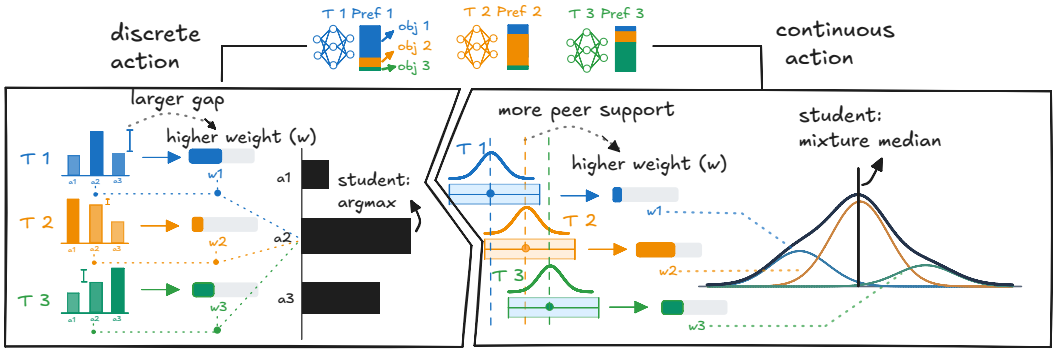


Figure 1: Overview of CAST. CAST aggregates smoothed teachers (T1 - T3), each with a different preference (Pref 1 - Pref 3), into a certified student policy. It assigns each teacher a weight based on its robustness and then perform a weighted aggregation of the teachers' policies. The robustness measure and policy representation are the action-value gap and Q-values (colored bars) of actions (a1, a2, and a3) in the discrete setting, and peer support and action distributions (colored curves) in the continuous setting.

- We evaluate CAST on five multi-objective DRL benchmarks. CAST preserves competitive target-preference utility under state perturbations, exceeds the best source teacher in 17 out of 40 attack configurations, and shows that teacher smoothing improves the student in 39 out of 40 configurations, with a maximum relative gain of 183.9%.

2 Related Work

CAST is related to certified robustness in DRL, zero-shot transfer learning in DRL, and mixture of experts in DRL. Certified robust DRL provides formal guarantees for policy behavior under bounded state perturbations [12, 15, 13, 20–27]. However, these methods certify a single policy under a fixed task, so their certificates cannot be directly reused when the task changes. Zero-shot transfer learning in DRL aims to improve policy generalization to unseen deployment environments without additional test-time training [28–38]. These methods mainly train a policy to generalize during training, whereas CAST enables training-free transfer without access to target-side data. Mixture of experts in DRL combines multiple expert components through gating, routing, or weighted aggregation [39–41, 7, 42–48]. Nevertheless, existing MoE-based DRL methods usually learn expert modules from training data, and they do not provide certified robustness after aggregating multiple pre-trained policies. Therefore, these lines of work do not address the training-free certified policy adaptation problem, where no target-side data are available. Detailed related work is in Appx. A.

3 Preliminaries

We introduce two preliminaries used in CAST: the multi-objective DRL setting for formulating the adaptation problem, and the smoothing-based certification.

Multi-Objective DRL. A multi-objective Markov Decision Process is defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \mathbf{r}, \gamma \rangle$, where $\mathbf{r}(s,a) \in \mathbb{R}^m$ is a vector reward. For a policy π, its vector-valued discounted cumulative reward is $\mathbf{J}(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t)]$. A preference vector $\omega \in \Delta^{m-1}$ defines a scalarized task by $r_\omega(s,a) := \omega^\top \mathbf{r}(s,a)$ and $J_\omega(\pi) := \omega^\top \mathbf{J}(\pi)$. Thus, different preferences can induce different optimal policies under the same transition dynamics. In CAST, source teachers are certified smoothed policies trained under different preferences, and the student is evaluated under a target preference $\omega^\star$.

Smoothing-Based Certified DRL. Smoothing-based certified DRL constructs a smoothed decision rule by evaluating a policy or value function under Gaussian perturbations of the input state. Since discrete and continuous action spaces require different certified objects, we introduce the two smoothing-based certification tools used in CAST separately.

For discrete action spaces, smoothing can be applied to an action-value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Given Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$, the smoothed action-value function is $\tilde{Q}(s, a) := \mathbb{E}_{\delta \sim \mathcal{N}(0,\sigma^2 I)}[Q(s + \delta, a)]$. The following Lem. 1, adapted from [12], provides an Lipschitz bound for smoothed action values. This bound allows us to estimate the worst-case change in each action value under a state perturbation of radius ϵ as $L_Q \epsilon$.

Lemma 1 (Lipschitzness of smoothed action values [12, Lem. 1])**.** *Assume that the base action-value function Q is bounded in $[V_{\min}, V_{\max}]$. Then, for every action $a \in \mathcal{A}$, the smoothed action-value function $\tilde{Q}(\cdot, a)$ is L_Q-Lipschitz in the state, i.e., $|\tilde{Q}(s, a) - \tilde{Q}(s', a)| \leq L_Q \|s - s'\|_2$, where $L_Q = \frac{V_{\max} - V_{\min}}{\sigma} \sqrt{\frac{2}{\pi}}$.*

For continuous action spaces, we follow the certified smoothing procedure of [13] and apply smoothing to the mean action output used for evaluation. Let $\mu_k(s)$ denote the k-th coordinate of this mean action. Given Gaussian input noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$, the p-quantile smoothed action coordinate is defined as

$$\tilde{\pi}_{k,p}(s) := \inf \left\{ a \in \mathbb{R} : \mathbb{P}_{\delta \sim \mathcal{N}(0,\sigma^2 I)} \left[\mu_k(s + \delta) \leq a \right] \geq p \right\}.$$

In our implementation, we set $p = 0.5$, in which case the quantile-smoothed action reduces to median smoothing. Lem. 2, adapted from [13], gives a certified coordinate-wise interval for quantile-smoothed continuous actions, allowing us to estimate the worst-case range of each action coordinate under a state perturbation of radius ϵ.

Lemma 2 (Action bound for quantile-smoothed continuous actions [13])**.** *Fix a state s, perturbation radius ϵ, smoothing scale σ, and quantile level $p \in (0, 1)$. Define $p_\epsilon^- = \Phi(\Phi^{-1}(p) - \epsilon/\sigma)$ and $p_\epsilon^+ = \Phi(\Phi^{-1}(p) + \epsilon/\sigma)$, where Φ is the standard Gaussian CDF. Then, for every s' satisfying $\|s' - s\|_2 \leq \epsilon$ and every action coordinate k, we have $\tilde{\pi}_{k,p_\epsilon^-}(s) \leq \tilde{\pi}_{k,p}(s') \leq \tilde{\pi}_{k,p_\epsilon^+}(s)$. Equivalently, the perturbed smoothed action lies in the coordinate-wise interval $[L_k(s, \epsilon), R_k(s, \epsilon)]$, where $L_k(s, \epsilon) := \tilde{\pi}_{k,p_\epsilon^-}(s)$ and $R_k(s, \epsilon) := \tilde{\pi}_{k,p_\epsilon^+}(s)$.*

4 CAST: Certifiable Aggregation of Smoothed Teachers

In this section, we present CAST, a robustness-aware aggregation framework for training-free policy adaptation. Given multiple certified smoothed teachers trained under different source preferences, CAST constructs a student policy for a target preference without target-side data, target-environment interaction, or additional retraining. The main challenge is that each teacher certificate only certifies an individual teacher decision under its own policy and preference, so it cannot be directly reused for the aggregated student policy. CAST addresses this challenge by extracting partial robustness information from teacher certificates and converting it into state-dependent aggregation weights, while still combining teacher decisions according to their target-preference utility. Since discrete and continuous action spaces have different certified objects, we instantiate CAST separately for the two settings below.

4.1 Aggregation for Discrete Action Spaces

For discrete action spaces, CAST uses each certified teacher's smoothed action Q-values and smoothed action margin. The smoothed Q-values preserve the teacher's preferences over all actions, while the margin measures the gap between the greedy action and its strongest competitor and serves as a local robustness signal. CAST uses this margin to build advantage-margin weights, aggregates the weighted smoothed Q-values into a student value function, and selects the student action greedily.

Aggregated student action. For each teacher $i \in [N]$, let Q_i denote its base action-value function, and let $\tilde{Q}_i$ denote the corresponding Gaussian-smoothed action-value function defined in Sec. 3. In the discrete setting, $\tilde{Q}_i(s, \cdot)$ assigns a smoothed value to every action in the finite action set. Given the smoothed teacher values $\{\tilde{Q}_i\}_{i=1}^N$ and a state-dependent teacher weight vector $w(s) = (w_1(s), \ldots, w_N(s)) \in \Delta^{N-1}$, CAST defines the aggregated student action-value function using a state-dependent log-sum-exp aggregation rule: $Q_{\text{agg}}(s, a) := \frac{1}{\lambda} \log\Big(\sum_{i=1}^N w_i(s) \exp(\lambda \tilde{Q}_i(s, a)) \Big)$, where $\lambda \neq 0$ controls the sharpness of the aggregation. The student policy is then obtained greedily: $\pi_S(s) := \arg\max_{a \in \mathcal{A}} Q_{\text{agg}}(s, a)$.

Advantage-margin weights. Lem. 1 shows that each Gaussian-smoothed action value has bounded variation under state perturbations. Based on this property, we use the gap between the largest smoothed action value and the second largest one as a robustness signal for each teacher. For teacher i, let $a_i^\star(s) := \arg\max_{a\in\mathcal{A}} \tilde{Q}_i(s,a)$ denote its smoothed greedy action at state s. We define the corresponding advantage margin as $m_i(s) := \tilde{Q}_i(s, a_i^\star(s)) - \max_{a\in\mathcal{A},\, a\neq a_i^\star(s)} \tilde{Q}_i(s,a)$. To obtain a local robustness score without introducing any student-side sampling, we compute this margin using the same Gaussian smoothing distribution used in the teacher certificate: $u_i(s) := \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[m_i(s+\delta)]$. The advantage-margin weights are then defined as: $w_i(s) := \frac{\exp(\beta u_i(s))}{\sum_{k=1}^N \exp(\beta u_k(s))}$, where $\beta > 0$ controls the sharpness of the weighting scheme. This construction gives larger influence to teachers whose smoothed greedy decisions have greater resistance to perturbations.

4.2 Aggregation for Continuous Action Spaces

For continuous action spaces, CAST uses each certified teacher's quantile-smoothed action vector and coordinate-wise certified action interval. The action vector provides the teacher proposal, and the interval provides the local robustness signal. CAST aggregates teacher proposals through a weighted mixture, defines the student action by a mixture quantile, and assigns larger weights to teachers whose actions better agree with other teachers' certified intervals.

Aggregated student action. For each certified teacher $i \in [N]$, let $a_i(s) := \tilde{\pi}_{i,p}(s) \in \mathbb{R}^{d_a}$ denote its quantile-smoothed action defined in Sec. 3, where $a_{i,k}(s)$ is the k-th action coordinate. Following [13], we set $p = 0.5$ throughout, so the quantile-smoothed action corresponds to the median action. Given a state-dependent teacher weight vector $w(s) = (w_1(s), \ldots, w_N(s)) \in \Delta^{N-1}$, CAST defines the student action coordinate-wise through a weighted Gaussian mixture centered at the teacher actions. For each coordinate $k \in [d_a]$, the student mixture CDF is $F_{\mathrm{stu},k}(a \mid s) := \sum_{i=1}^N w_i(s)\, \Phi\Big(\frac{a - a_{i,k}(s)}{\sigma_a}\Big)$, where Φ is the standard Gaussian CDF and $\sigma_a > 0$ is a fixed action-space smoothing scale. The k-th coordinate of the student action is defined as the p-quantile of this mixture: $a_{\mathrm{stu},k}(s) := \inf\{a \in \mathbb{R} : F_{\mathrm{stu},k}(a \mid s) \geq p\}$, which is computed using Brent's method [49] in our implementation. The continuous student policy is then $\pi_S(s) := a_{\mathrm{stu}}(s) = (a_{\mathrm{stu},1}(s), \ldots, a_{\mathrm{stu},d_a}(s))$.

Peer-review weights. We next define the state-dependent mixture weights. By Lem. 2, each teacher j provides a coordinate-wise certified action interval $[L_{j,k}(x,\epsilon), R_{j,k}(x,\epsilon)]$, $k \in [d_a]$, under state perturbations. These intervals allow CAST to measure whether one teacher's action is supported by the certified action regions of other teachers. For $z \in \mathbb{R}^{d_a}$, define the hinge distance from z to teacher j's certified interval as $\mathrm{dist}\big(z, [L_j(x,\epsilon), R_j(x,\epsilon)]\big) := \sum_{k=1}^{d_a} [\max\{0, L_{j,k}(x,\epsilon) - z_k\} + \max\{0, z_k - R_{j,k}(x,\epsilon)\}]$. The peer-review loss of teacher i is $\mathcal{L}_i(x) := \sum_{j\neq i} \mathrm{dist}\big(a_i(x), [L_j(x,\epsilon), R_j(x,\epsilon)]\big)$. A smaller loss indicates that teacher i's action is more consistent with its peers' certified intervals. CAST then selects the most consistent teacher by $f(x) := \arg\min_{i\in[N]} \mathcal{L}_i(x)$, and defines the peer-review weight of teacher i as its winning probability under Gaussian smoothing: $w_i(s) := \mathbb{P}[f(s+\delta) = i], \delta \sim \mathcal{N}(0, \sigma^2 I)$.

5 Robustness Certification for CAST

In this section, we discuss three robustness certification of CAST: the discrete action bound, the continuous action bound, and a unified cumulative reward certificate for both discrete and continuous action spaces.

5.1 Action Certification

Theorem 1 (Discrete action certification)**.** *For every fixed action $a \in \mathcal{A}$, the aggregated value $s \mapsto Q_{\mathrm{agg}}(s,a)$ is L_{agg}-Lipschitz with $L_{\mathrm{agg}} := L_Q + \frac{4\beta L_Q}{|\lambda|}$. Let $a^\star_{\mathrm{agg}}(s) \in \arg\max_{a\in\mathcal{A}} Q_{\mathrm{agg}}(s,a)$ be the greedy action of the aggregated student, and define the aggregated action gap as $g_{\mathrm{agg}}(s) := Q_{\mathrm{agg}}(s, a^\star_{\mathrm{agg}}(s)) - \max_{a\in\mathcal{A},\, a\neq a^\star_{\mathrm{agg}}(s)} Q_{\mathrm{agg}}(s,a)$. If $g_{\mathrm{agg}}(s) > 0$, then the student action is certified to remain unchanged for every perturbed state s' satisfying $\|s' - s\|_2 < r_{\mathrm{dis}}(s) := \frac{g_{\mathrm{agg}}(s)}{2L_{\mathrm{agg}}}$. That is, $\pi_S(s') = \pi_S(s)$.*

Theorem 2 (Continuous action certification)**.** *Fix a state s, perturbation radius ϵ, quantile level $p \in (0,1)$, and action coordinate $k \in [d_a]$. Suppose that, for every teacher $i \in [N]$, Lem. 2 gives certified bounds $L_{i,k}(s,\epsilon)$ and $R_{i,k}(s,\epsilon)$ such that $a_{i,k}(s') \in [L_{i,k}(s,\epsilon), R_{i,k}(s,\epsilon)]$, for all $\|s'-s\|_2 \le \epsilon$. Define $d_{i,k}(s,\epsilon) := R_{i,k}(s,\epsilon) - L_{i,k}(s,\epsilon)$ and $\Delta_k(s,\epsilon) := 2\left(2\Phi\left(\frac{\epsilon}{2\sigma}\right)-1\right) + \max_{i\in[N]}\left(2\Phi\left(\frac{d_{i,k}(s,\epsilon)}{2\sigma_a}\right)-1\right)$. Assume $\Delta_k(s,\epsilon) < \min\{p, 1-p\}$. Let $a_{\mathrm{stu},k,q}(s)$ denote the q-quantile of the clean student mixture CDF $F_{\mathrm{stu},k}(\cdot \mid s)$. Then for every perturbed state s' satisfying $\|s'-s\|_2 \le \epsilon$, $a_{\mathrm{stu},k,p-\Delta_k(s,\epsilon)}(s) \le a_{\mathrm{stu},k,p}(s') \le a_{\mathrm{stu},k,p+\Delta_k(s,\epsilon)}(s)$.*

Thms. 1 and 2 provide action certificates for the discrete- and continuous-action spaces of CAST. At a high level, their proofs follow Thm. 1 in [12] and Thm. 3 in [13], respectively. Compared with certification of [12] and [13], certifying CAST additionally requires bounding the effect of state perturbations on the aggregation weights. Intuitively, this is necessary as the student's policy from CAST depends on both the actions of the smoothed teachers and their state-dependent aggregation weights. The state perturbation can therefore affect both the teachers' actions and their weights on student policies.

Since CAST uses different aggregations in the two action settings, bounding the weight variation under perturbation is different. In the discrete-action setting, the variation of the advantage-margin weights is controlled by $|\log w_i(s') - \log w_i(s)| \le 4\beta L_Q \|s'-s\|_2$, which contributes the additional term $4\beta L_Q/|\lambda|$ to L_{agg}. In the continuous-action setting, the variation of the peer-review weights is bounded by $\|w(s') - w(s)\|_1 \le 2\left(2\Phi\left(\frac{\epsilon}{2\sigma}\right)-1\right)$, which forms the first term in $\Delta_k(s,\epsilon)$. The full proofs are in Appx. B and C, respectively.

5.2 Cumulative Reward Certification

To derive a cumulative reward certificate, we combine the certified action set in Def. 1 with the environment smoothness condition in Assump. 1. These two ingredients bound how state perturbations and possible action changes affect the student's cumulative reward, leading to Thm. 3.

Definition 1 (Certified action set)**.** Fix a reference state s and radius $\epsilon \ge 0$. A set $C(s,\epsilon) \subseteq \mathcal{A}$ is a certified action set for the student policy π_S if $\pi_S(s') \in C(s,\epsilon), \forall s'$ such that $d_{\mathcal{S}}(s',s) \le \epsilon$.

Assumption 1 (Vector valued environment smoothness)**.** *We use the state metric $d_{\mathcal{S}}(s,\hat{s}) = \|s-\hat{s}\|_2$. For discrete action spaces, let $d_{\mathcal{A}}(a,\hat{a}) = \mathbf{1}[a \ne \hat{a}]$. For continuous action spaces with box constraints $a \in [a_{\min}, a_{\max}] \subset \mathbb{R}^{d_a}$, let $d_{\mathcal{A}}(a,\hat{a}) = \|(a-\hat{a}) \oslash (a_{\max}-a_{\min})\|_2$, where $\oslash$ denotes element-wise division. Define the joint metric as $d_{\mathcal{SA}}((s,a),(\hat{s},\hat{a})) := d_{\mathcal{S}}(s,\hat{s}) + d_{\mathcal{A}}(a,\hat{a})$. There exist constants $L_P^{\mathrm{TV}}, L_R^{\mathrm{vec}} > 0$ and $R_{\max}^{\mathrm{vec}} > 0$ such that $\|\mathbf{r}(s,a)\|_\infty \le R_{\max}^{\mathrm{vec}}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and for all $(s,a),(\hat{s},\hat{a}) \in \mathcal{S}\times\mathcal{A}$, $\mathrm{TV}(P(\cdot \mid s,a), P(\cdot \mid \hat{s},\hat{a})) \le L_P^{\mathrm{TV}}\, d_{\mathcal{SA}}((s,a),(\hat{s},\hat{a}))$, $\|\mathbf{r}(s,a) - \mathbf{r}(\hat{s},\hat{a})\|_\infty \le L_R^{\mathrm{vec}}\, d_{\mathcal{SA}}((s,a),(\hat{s},\hat{a}))$.*

Theorem 3 (Vector-valued cumulative reward certificate)**.** *Assume an infinite-horizon discounted MDP with discount factor $\gamma \in [0,1)$. Suppose that the CAST student admits a certified action set $C(s,\epsilon)$. Define $\Delta_A(s,\epsilon) := \sup_{a,b\in C(s,\epsilon)} d_{\mathcal{A}}(a,b)$, $V_{\max}^{\mathrm{vec}} := \frac{R_{\max}^{\mathrm{vec}}}{1-\gamma}$. Under Assump. 1, for any s' satisfying $d_{\mathcal{S}}(s',s) \le \epsilon$, $\|\mathbf{V}^{\pi_S}(s') - \mathbf{V}^{\pi_S}(s)\|_\infty \le \left(L_R^{\mathrm{vec}} + 2\gamma V_{\max}^{\mathrm{vec}} L_P^{\mathrm{TV}}\right)(\epsilon + \Delta_A(s,\epsilon))$.*

The proof of Thm. 3 connects the action certificate to the cumulative reward certificate. Intuitively, a state perturbation may change the student's action, which can further affect its cumulative reward. To formalize this intuition, the certified action set $C(s,\epsilon)$ ensures that both the clean action $\pi_S(s)$ and the perturbed action $\pi_S(s')$ belong to the same certified set, so that their action distance is bounded by $\Delta_A(s,\epsilon)$. Given the action distance bound and the fact that the state variation is bounded by ϵ, the joint state-action variation is bounded by $\epsilon + \Delta_A(s,\epsilon)$. Under Assump. 1, applying the Bellman equation then yields a cumulative reward bound. The full proof is in Appx. D.

6 Experiments

Environment and preference setup. We evaluate CAST[3] on five multi-objective reinforcement learning environments from MO-Gymnasium [50]. For continuous control, we use `mo-hopper-v4`, `mo-walker2d-v4`, and `mo-humanoid-v4`, following the multi-objective MuJoCo setting used in

[3]CAST source code: `https://github.com/Ryndalf/CAST`.

[13] and [51]. For discrete control, we follow Chen et al. [51] and use `deep-sea-treasure-v0` and `fruit-tree-v0`. Since all tasks are multi-objective, each policy is associated with a reward preference vector. We evaluate the student under the average target preference $\omega^\star = (1/d, \ldots, 1/d)$ for a task with d reward dimensions, and select source teachers whose preferences emphasize different reward dimensions. Teacher identifiers encode their preference weights after removing decimal points; for example, `teacher-19` denotes $(0.1, 0.9)$.

Backbones and Baselines. We use two smoothing-based certified DRL methods as certification backbones for constructing the certified teacher models. For continuous-control tasks, we borrow the quantile-smoothing certification technique from [13]. For discrete-action tasks, we build on CROP [12]. These backbones are used to smooth and certify the teacher policies. To evaluate CAST against an existing method that most closely matches our setting, we include MORL-FB [51] as a zero-shot transfer baseline. The detailed rationale behind these baseline choices is in Appx. E.

Attacker models. We consider ℓ_2-bounded adversarial state perturbations. Under this threat model, we evaluate CAST using a set of representative state attacks commonly used in robust DRL [52, 13, 53]. Specifically, for continuous-control tasks, we evaluate random perturbations, MAD, SARSA [9], and PAAD attacks [54]. For discrete-action tasks, we evaluate random perturbations, FGSM [55] and PGD [56], and PAAD attacks [54].

6.1 Experimental Questions

Our experiments answer four questions: (1) whether CAST preserves target-preference utility under changed preferences; (2) whether smoothing improves robustness against state attacks; (3) under what conditions the derived certificates remain non-trivial, i.e., yielding a positive certified radius in discrete action spaces or a certified interval narrower than the full action range in continuous action spaces; and (4) whether the actions and cumulative rewards remain within their corresponding certified bounds under attacks. The full parameter selection protocol, parameter-selection tables, runtime, computational resources, and teacher or attacker training details are in Appx. F and Appx. G.

Student-Preference Utility under State Perturbations. Tab. 1 reports representative student preference utility results on Walker2d and Deep Sea Treasure. On both tasks, the CAST student achieves competitive performance compared with the certified teachers and often matches or exceeds the best teacher under state perturbations. These results suggest that CAST can preserve useful target preference behavior when aggregating teachers trained under different source preferences. Across the full results in Appx. H, CAST exceeds the best teacher in seventeen out of forty attack configurations, including thirteen out of twenty-four continuous control configurations and four out of sixteen discrete control configurations.

Since aggregating teachers with different source preferences may reduce an individual teacher's advantage against a particular attack, CAST is not expected to outperform the best teacher in every configuration. That is, achieving competitive target-preference utility without any target-side data is challenging. For example, the comparison with MORL-FB highlights the significance of this result. Despite using target-side data and interacting with the target environment during training, MORL-FB fails to outperform the lowest-reward teacher in fourteen out of the sixteen configurations. By contrast, CAST never falls below the lowest-reward teacher in these configurations, outperforms MORL-FB in fifteen out of the sixteen, and ties it in one. These results demonstrate that CAST can preserve useful teacher knowledge, largely avoid negative transfer, and adapt to a new target preference without target-side information.

To further examine whether CAST preserves target-preference utility beyond the balanced target preference, we also test additional student preferences in Appx. I. This sensitivity study shows that CAST is not limited to the average preference: across the extended preference comparisons, the student remains within the teacher reward range in 57 out of 116 cases and exceeds the best teacher in 44 out of 116 cases. Thus, CAST either stays within the teacher range or improves over all teachers in 101 out of 116 cases under changed target preferences.

Effect of Smoothing on Student Robustness. Tab. 2 reports representative smoothing comparison results on Walker2d and Deep Sea Treasure. On both tasks, smoothed student aggregation generally achieves higher attacked rewards than direct student aggregation without smoothing, showing that the robustness benefit of certified teachers can be transferred to the CAST student. Across the full results in Appx. J, smoothing improves the attacked return in thirty-nine of forty configurations, with

Table 1: Representative utility comparison among teachers, MORL-FB [51], and the smoothed student on Walker2d and Deep Sea Treasure under attacks (bold values indicate the best reward in each comparison group, and underlined values indicate the second-best reward).

Walker2d ($\sigma = 0.1$)					
ϵ	Model	Rand	MAD	SARSA	PAAD
0.3	teacher-91	$\underline{2124.72 \pm 64.78}$	$\underline{1261.10 \pm 502.47}$	842.02 ± 296.14	$\mathbf{1681.05 \pm 331.92}$
	teacher-19	993.12 ± 0.61	992.55 ± 0.68	$\underline{989.30 \pm 0.56}$	990.70 ± 1.02
	MORL-FB	1791.79 ± 600.77	-11.11 ± 93.01	553.44 ± 868.94	1113.00 ± 830.85
	stu. (CAST)	$\mathbf{2368.93 \pm 18.45}$	$\mathbf{2003.55 \pm 397.72}$	$\mathbf{1632.13 \pm 834.98}$	$\underline{1263.13 \pm 778.95}$
1.0	teacher-91	$\mathbf{1263.68 \pm 447.34}$	-54.86 ± 54.72	-52.17 ± 27.83	64.48 ± 246.92
	teacher-19	985.60 ± 1.28	$\mathbf{437.32 \pm 299.79}$	$\mathbf{499.68 \pm 303.86}$	$\underline{115.47 \pm 20.47}$
	MORL-FB	-2.28 ± 48.55	-270.80 ± 107.11	-102.69 ± 16.58	-36.72 ± 28.51
	stu. (CAST)	$\underline{1096.41 \pm 719.40}$	$\underline{117.67 \pm 33.49}$	$\underline{93.63 \pm 31.35}$	$\mathbf{160.41 \pm 101.00}$
Deep Sea Treasure ($\sigma = 0.5$)					
ϵ	Model	Rand	FGSM	PGD	PAAD
0.70	teacher-37	$\mathbf{2.46 \pm 0.91}$	$\mathbf{2.85 \pm 1.02}$	$\underline{2.23 \pm 1.56}$	$\mathbf{2.93 \pm 0.33}$
	teacher-73	2.35 ± 0.00	$\underline{2.35 \pm 0.00}$	$\mathbf{2.35 \pm 0.00}$	$\underline{2.35 \pm 0.00}$
	MORL-FB	-5.35 ± 12.88	-5.48 ± 12.82	-2.35 ± 11.33	-10.76 ± 14.24
	stu. (CAST)	$\underline{2.39 \pm 0.11}$	$\underline{2.35 \pm 0.00}$	$\mathbf{2.35 \pm 0.00}$	$\underline{2.35 \pm 0.00}$
3.00	teacher-37	$\underline{1.60 \pm 1.33}$	-0.15 ± 0.00	-0.15 ± 0.00	$\underline{-25.00 \pm 0.00}$
	teacher-73	1.11 ± 1.67	$\underline{2.67 \pm 0.20}$	$\underline{2.60 \pm 0.00}$	$\mathbf{2.39 \pm 0.11}$
	MORL-FB	-4.36 ± 11.51	-6.56 ± 12.16	-2.95 ± 11.07	$\underline{-25.00 \pm 0.00}$
	stu. (CAST)	$\mathbf{2.01 \pm 1.12}$	$\mathbf{2.93 \pm 0.33}$	$\mathbf{2.89 \pm 0.37}$	$\underline{-25.00 \pm 0.00}$

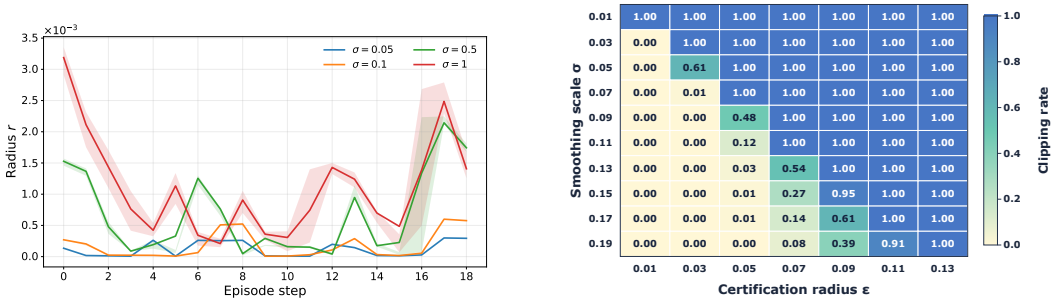
Table 2: Representative smoothing comparison between direct student aggregation without smoothing and smoothed student aggregation on Walker2d ($\sigma = 0.1$) and Deep Sea Treasure ($\sigma = 0.5$) under attacks. Higher reward indicates stronger robustness under attack.

Walker2d ($\sigma = 0.1$)					
ϵ	Model	Rand	MAD	SARSA	PAAD
0.3	student (w/o smoothing)	834.43 ± 341.97	682.41 ± 441.58	527.46 ± 547.19	840.21 ± 395.91
	student (w/ smoothing)	2368.93 ± 18.45	2003.55 ± 397.72	1632.13 ± 834.98	1263.13 ± 778.95
1.0	student (w/o smoothing)	448.17 ± 288.63	-16.21 ± 124.91	-10.38 ± 119.19	-63.68 ± 43.92
	student (w/ smoothing)	1096.41 ± 719.40	117.67 ± 33.49	93.63 ± 31.35	160.41 ± 101.00
Deep Sea Treasure ($\sigma = 0.5$)					
ϵ	Model	Rand	FGSM	PGD	PAAD
0.70	student (w/o smoothing)	1.37 ± 2.26	-5.67 ± 12.67	-8.31 ± 13.64	-14.06 ± 13.40
	student (w/ smoothing)	2.39 ± 0.11	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00
3.00	student (w/o smoothing)	1.96 ± 1.60	-0.58 ± 8.23	-2.79 ± 11.11	-22.52 ± 7.46
	student (w/ smoothing)	2.01 ± 1.12	2.93 ± 0.33	2.89 ± 0.37	-25.00 ± 0.00

a maximum relative gain of 183.9%. This overall improvement suggests that CAST can transfer the robustness benefit of teacher smoothing to the aggregated student under state perturbations.

Sensitivity of Action Certification. In order to understand when CAST provides non-trivial action guarantees, we examine the sensitivity of the discrete- and continuous-action certificates. Since the two certificates take different forms according to Thms. 1 and 2, we analyze how σ affects the certified radius in the discrete-action setting and how σ relative to the perturbation radius ϵ affects the clipping rate in the continuous-action setting, where the clipping rate measures how often either endpoint $p - \Delta_k(s, \epsilon)$ or $p + \Delta_k(s, \epsilon)$ reaches the boundary of $[0, 1]$.

For the discrete action setting, Fig. 2a reports the certified radius across different state smoothing scales σ via empirically evaluating the discrete action certificate in Thm. 1. At smoothing scales $\sigma \in \{0.05, 0.1\}$, several evaluated steps have a certified radius of zero, meaning that, with smoothing



(a) Deep Sea Treasure certified radius

(b) Walker2d clipping rate

Figure 2: Representative action-level certificate diagnostics for discrete and continuous action spaces.

scales that are small relative to $\sigma \in \{0.5, 1.0\}$, the certificate cannot guarantee that the student action remains unchanged under any nonzero state perturbation. As σ increases, the certified radius generally becomes larger and remains nonzero more often, showing that the larger smoothing scale is needed to provide non-trivial action robustness guarantees. Results for the other environments are in Appx. K.

For continuous action setting, we introduce the clipping rate to assess whether the certified action interval obtained from Thm. 2 is non-trivial, i.e., narrower than full action range. In Thm. 2, the lower and upper action bounds of certified action interval are given by the quantiles of the student mixture distribution at levels $p - \Delta_k(s, \epsilon)$ and $p + \Delta_k(s, \epsilon)$, respectively. When either quantile level exceeds the boundaries of $[0, 1]$, it is clipped to 0 or 1. With $p = 0.5$ in our setting, both levels are clipped simultaneously to 0 and 1, yielding a certified action interval that covers the full action range, resulting in a trivial certificate for the student action. The clipping rate is the proportion of evaluated cases across random seeds in which the quantile level is clipped, and therefore reflects, to some extent, the non-triviality of the certificate from Thm. 2.

Fig. 2b illustrates when the certified action intervals are non-trivial based on the clipping rate. Obtaining non-trivial certified action intervals in some but not all evaluated cases, i.e., clipping rates strictly between zero and one, requires a ratio σ/ϵ of at least 1.66 (observed at $(\sigma, \epsilon) = (0.05, 0.03)$). Obtaining non-trivial certified action intervals in all evaluated cases, i.e., a zero clipping rate, requires a ratio of at least 3.0 (observed at $(\sigma, \epsilon) = (0.03, 0.01)$). Therefore, non-trivial certified action intervals are observed only when the ratio σ/ϵ is at least 1.66 in Walker2d. The ratio can vary across environments and teacher policies, as the second term in $\Delta_k(s, \epsilon)$ depends on $d_{i,k}(s, \epsilon)$, the widths of the teachers' certified action intervals in the corresponding environment. More detailed results are in Appx. K.

Empirical Evaluation of Certification Guarantees. In order to empirically evaluate CAST's certification guarantees, we validate its action certificates and cumulative reward lower bounds. Based on the results reported in **Sensitivity of Action Certification**, we use the largest evaluated smoothing scales, i.e., $\sigma = 1.0$ for Deep Sea Treasure and $\sigma = 0.19$ for Walker2d, to obtain certificates for which the discrete certified radius is nonzero and the continuous clipping rate is below one. Tab. 3 reports the empirical evaluation of the resulting action certificates.

For the discrete action setting, we select ten fixed states from Deep Sea Treasure. At each state, we compute the certified radius and report its mean over the ten states. We then apply four attacks under six perturbation magnitudes and report the percentage of attacked samples whose student action remains identical to the clean action. The mean certified radius is 1.32×10^{-3}. All attacked samples whose actual perturbation norms fall within this radii preserve the clean action, empirically supporting the action guarantee. Moreover, all four attacks achieve a 100% action preservation rate even at $\epsilon = 10^{-2}$ and $\epsilon = 10^{-1}$, which are larger than the mean certified radius. Action changes only begin to appear at $\epsilon = 1$ and become more frequent at $\epsilon = 10$. These results suggest that **the certificate is valid, and in practice, the student action can remain stable well beyond the theoretical guarantee**. For the continuous action setting, we select ten fixed states from Walker2d. For each state and perturbation magnitude, we compute the coordinate-wise certified action interval and evaluate four attacks with ten initializations per state. We report the minimum certified interval

Table 3: Representative empirical evaluation of the action-level certificates.

Discrete action: Deep Sea Treasure ($\sigma = 1.0$, mean certified radius $= 1.32\times10^{-3}$)							
Environment	Attack	$\epsilon = 10^{-4}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-1}$	$\epsilon = 1$	$\epsilon = 10$
Deep Sea Treasure	Rand	100%	100%	100%	100%	88%	62%
	FGSM	100%	100%	100%	100%	40%	10%
	PGD	100%	100%	100%	100%	41%	31%
	PAAD	100%	100%	100%	100%	62%	35%

Continuous action: Walker2d ($\sigma = 0.19$)							
Environment	ϵ	Minimum certified interval diagonal (L_2)	Maximum action L_2 shift				Coverage across all attacks
			Rand	MAD	SARSA	PAAD	
Walker2d	10^{-4}	0.006355	0.000163	0.003719	0.000229	0.000301	100.00%
	10^{-3}	0.063547	0.008538	0.008388	0.005093	0.008288	100.00%
	10^{-2}	0.688459	0.021666	0.037412	0.030383	0.027429	100.00%
	10^{-1}	4.891142	0.133555	0.343670	0.175636	0.302095	100.00%
	1	4.898980	1.453948	1.919273	1.725271	1.632178	100.00%
	10	4.898980	2.050822	3.067277	2.615040	2.990106	100.00%

width across the ten states, the maximum observed action shift under each attack, and the coverage rate, which measures the percentage of attacked actions contained within the certified bounds. The certified intervals achieve 100% coverage under all evaluated perturbation magnitudes and attacks, showing that all observed attacked actions lie within the certified ranges. However, the interval width increases with ϵ. At $\epsilon = 10^{-1}$, the certified interval already approaches the full action range, and when $\epsilon \geq 1$, it covers the entire action range and thus becomes vacuous. Complete results for all environments are in Appx. L.

For the cumulative reward certificate, we report representative results on Walker2d and Deep Sea Treasure. On Walker2d, the worst empirical rewards are $1,263.13$ and 93.63 at $\epsilon = 0.3$ and $\epsilon = 1.0$, respectively, both remaining above their certified lower bounds of $-4,100.13$ and $-6,416.25$. Similarly, on Deep Sea Treasure, the worst empirical rewards of 2.35 and -25.00 at $\epsilon = 0.7$ and $\epsilon = 3.0$ remain above the corresponding lower bounds of $-1,690.15$ and $-3,980.06$. No empirical reward violates its certified lower bound. Complete results are in Appx. M.

6.2 Ablation Study

We conduct two ablation studies to evaluate CAST's aggregation design: (1) examining the contributions of CAST's teacher weights and nonlinear aggregation by comparing equal-weight linear aggregation, CAST-weighted linear aggregation, equal-weight nonlinear aggregation, and full CAST with both components enabled; and (2) comparing CAST with simple baselines, including selecting a single teacher uniformly or according to CAST's teacher weights and using its action as the student action, as well as aggregating teacher policies with equal weights. The results show that the complete CAST aggregation achieves better overall performance than either component alone and the simple aggregation methods, validating the effectiveness of the proposed design. Detailed results are in Appx. N.

7 Conclusion

We introduced CAST, a certifiable aggregation framework for training-free robust policy adaptation from multiple smoothed teachers to a target-preference student. CAST aggregates certified teacher outputs and derives action-level certificates for both discrete and continuous action spaces. We further instantiate a scalarized cumulative reward certificate for multi-objective DRL. Experiments on five benchmarks show that CAST preserves competitive target-preference utility under state perturbations.

Acknowledgments and Disclosure of Funding

Zhongrui Zhao is supported by James Cook University Postgraduate Research Scholarship (JCUPRS). Zhongrui Zhao and Yanan Cai are supported by JCU Fee Sponsorship for International HDR Candidates.

References

[1] Shunyu Tian, Changyun Wei, Shaojie Jian, and Ze Ji. Preference-based deep reinforcement learning with automatic curriculum learning for map-free UGV navigation in factory-like environments. *Engineering Science and Technology, an International Journal*, 70:102147, 2025.

[2] Chien-Lun Cheng, Chen-Chien Hsu, Saeed Saeedvand, and Jun-Hyung Jo. Multi-objective crowd-aware robot navigation system using deep reinforcement learning. *Applied Soft Computing*, 151:111154, 2024.

[3] Guangran Cheng, Yuanda Wang, Lu Dong, Wenzhe Cai, and Changyin Sun. Multi-objective deep reinforcement learning for crowd-aware robot navigation with dynamic human preference. *Neural Computing and Applications*, 35(22):16247–16265, 2023.

[4] Samuel Garcin, James Doran, Shangmin Guo, Christopher G. Lucas, and Stefano V. Albrecht. DRED: Zero-shot transfer in reinforcement learning via data-regularised environment design. In *International Conference on Machine Learning*, pages 15006–15036, 2024.

[5] Shuai Zhang, Heshan Devaka Fernando, Miao Liu, Keerthiram Murugesan, Songtao Lu, Pin-Yu Chen, Tianyi Chen, and Meng Wang. SF-DQN: Provable knowledge transfer using successor feature for deep reinforcement learning. In *International Conference on Machine Learning*, pages 58897–58934, 2024.

[6] Guy Azran, Mohamad H. Danesh, Stefano V. Albrecht, and Sarah Keren. Contextual pre-planning on reward machine abstractions for enhanced transfer in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 10953–10961, 2024.

[7] Johan Samir Obando Ceron, Ghada Sokar, Timon Willi, Clare Lyle, Jesse Farebrother, Jakob Nicolaus Foerster, Gintare Karolina Dziugaite, Doina Precup, and Pablo Samuel Castro. Mixtures of experts unlock parameter scaling for deep RL. In *International Conference on Machine Learning*, pages 38520–38540, 2024.

[8] Ziye Huang, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Efficient residual learning with mixture-of-experts for universal dexterous grasping. In *International Conference on Learning Representations*, 2025.

[9] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems*, volume 33, pages 21024–21037, 2020.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[11] Pratik Vaishnavi, Kevin Eykholt, and Amir Rahmati. A study of the effects of transfer learning on adversarial robustness. *Transactions on Machine Learning Research*, 2024.

[12] Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations*, 2022.

[13] Chung-En Sun, Sicun Gao, and Tsui-Wei Weng. Breaking the barrier: Enhanced utility and robustness in smoothed DRL agents. In *International Conference on Machine Learning*, pages 46957–46987, 2024.

[14] Kiarash Banihashem, Adish Singla, and Goran Radanovic. Defense against reward poisoning attacks in reinforcement learning. *Transactions on Machine Learning Research*, 2023.

[15] Ronghui Mu, Leandro Soriano Marcolino, Yanghao Zhang, Tianle Zhang, Xiaowei Huang, and Wenjie Ruan. Reward certification for policy smoothed reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 21429–21437, 2024.

[16] Zhangheng Li, Tianlong Chen, Linyi Li, Bo Li, and Zhangyang Wang. Sparse transfer learning accelerates and enhances certified robustness: A comprehensive study. In *AAAI Conference on Artificial Intelligence*, volume 39, pages 5084–5091, 2025.

[17] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320, 2019.

[18] Yan Scholten, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. Hierarchical randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 36, pages 49783–49813, 2023.

[19] Motasem Alfarra, Adel Bibi, Philip H. S. Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pages 64–74, 2022.

[20] Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. In *International Conference on Learning Representations*, 2022.

[21] Xiaolin Sun and Zizhan Zheng. Belief-enriched pessimistic Q-learning against adversarial state perturbations. In *International Conference on Learning Representations*, 2024.

[22] Buqing Nie, Jingtian Ji, Yangqing Fu, and Yue Gao. Improve robustness of reinforcement learning against observation perturbations via ℓ_∞ Lipschitz policy networks. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 14457–14465, 2024.

[23] Chenxi Yang, Greg Anderson, and Swarat Chaudhuri. Certifiably robust reinforcement learning through model-based abstract interpretation. In *IEEE Conference on Secure and Trustworthy Machine Learning*, pages 233–251. IEEE, 2024.

[24] Chuangchuang Sun, Dong-Ki Kim, and Jonathan P. How. ROMAX: Certifiably robust deep multiagent reinforcement learning via convex relaxation. In *International Conference on Robotics and Automation*, pages 5503–5510. IEEE, 2022.

[25] Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, Gaojie Jin, and Qiang Ni. Certified policy smoothing for cooperative multi-agent reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 15046–15054, 2023.

[26] Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. COPA: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *International Conference on Learning Representations*, 2022.

[27] Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust policy learning against adversarial multi-agent communication. In *International Conference on Learning Representations*, 2023.

[28] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76: 201–264, 2023.

[29] Gabriele Tiboni, Pascal Klink, Jan Peters, Tatiana Tommasi, Carlo D'Eramo, and Georgia Chalvatzaki. Domain randomization via entropy maximization. In *International Conference on Learning Representations*, 2024.

[30] Abdus Salam Azad, Izzeddin Gur, Jasper Emhoff, Nathaniel Alexis, Aleksandra Faust, Pieter Abbeel, and Ion Stoica. CLUTR: Curriculum learning via unsupervised task representation learning. In *International Conference on Machine Learning*, pages 1361–1395, 2023.

[31] Yang Li, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Cooperative open-ended learning framework for zero-shot coordination. In *International Conference on Machine Learning*, pages 20470–20484, 2023.

[32] Wonil Song, Hyesong Choi, Kwanghoon Sohn, and Dongbo Min. A simple framework for generalization in visual RL under dynamic scene perturbations. In *Advances in Neural Information Processing Systems*, volume 37, pages 121790–121826, 2024.

[33] Di Zhang, Bowen Lv, Hai Zhang, Feifan Yang, Junqiao Zhao, Hang Yu, Chang Huang, Hongtu Zhou, Chen Ye, and Changjun Jiang. Focus on what matters: Separated models for visual-based RL generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 116960–116986, 2024.

[34] Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 26532–26548, 2022.

[35] Beyazit Yalcinkaya, Niklas Lauffer, Marcell Vazquez-Chanlatte, and Sanjit A. Seshia. Compositional automata embeddings for goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 72933–72963, 2024.

[36] Xinyue Wang and Biwei Huang. Modeling unseen environments with language-guided composable causal components in reinforcement learning. In *International Conference on Learning Representations*, 2025.

[37] Rui Yang, Jie Wang, Qijie Peng, Ruibo Guo, Guoping Wu, and Bin Li. Learning robust representations with long-term information for generalization in visual reinforcement learning. In *International Conference on Learning Representations*, 2025.

[38] James Queeney, Xiaoyi Cai, Alexander Schperberg, Radu Corcodel, Mouhacine Benosman, and Jonathan P. How. GRAM: Generalization in deep RL with a robust adaptation module. *IEEE Robotics and Automation Letters*, 11(2):2218–2225, 2026.

[39] Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

[40] Timon Willi, Johan Samir Obando Ceron, Jakob Nicolaus Foerster, Gintare Karolina Dziugaite, and Pablo Samuel Castro. Mixture of experts in a mixture of RL settings. *Reinforcement Learning Journal*, 3:1072–1105, 2024.

[41] Yilun Kong, Guozheng Ma, Qi Zhao, Haoyu Wang, Li Shen, Xueqian Wang, and Dacheng Tao. Mastering massive multi-task reinforcement learning via mixture-of-expert decision transformer. In *International Conference on Machine Learning*, pages 31379–31396, 2025.

[42] Wenhao Wu, Fuhong Liu, Haoru Li, Zican Hu, Daoyi Dong, Chunlin Chen, and Zhi Wang. Mixture-of-experts meets in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pages 24751–24785, 2025.

[43] Ghada Sokar, Johan Samir Obando Ceron, Aaron Courville, Hugo Larochelle, and Pablo Samuel Castro. Don't flatten, tokenize! Unlocking the key to SoftMoE's efficacy in deep RL. In *International Conference on Learning Representations*, 2025.

[44] Ahmed Hendawy, Jan Peters, and Carlo D'Eramo. Multi-task reinforcement learning with mixture of orthogonal experts. In *International Conference on Learning Representations*, 2024.

[45] Suning Huang, Zheyu Aqa Zhang, Tianhai Liang, Yihan Xu, Zhehao Kou, Chenhao Lu, Guowei Xu, Zhengrong Xue, and Huazhe Xu. MENTOR: Mixture-of-experts network with task-oriented perturbation for visual reinforcement learning. In *International Conference on Machine Learning*, pages 26143–26161, 2025.

[46] Wenxuan Song, Han Zhao, Pengxiang Ding, Can Cui, Shangke Lyu, Yaning Fan, and Donglin Wang. GeRM: A generalist robotic model with mixture-of-experts for quadruped robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 11879–11886. IEEE, 2024.

[47] Shengtian Yang, Yu Li, Shuo He, Yewen Li, Qingpeng Cai, Peng Jiang, and Lei Feng. Phase-aware mixture of experts for agentic reinforcement learning. In *International Conference on Machine Learning*, pages 145109–145133, 2026.

[48] Dohwan Ko, Jinyoung Park, Seoung Choi, Sanghyeok Lee, Seohyun Lee, and Hyunwoo J. Kim. MoE-GRPO: Optimizing mixture-of-experts via reinforcement learning in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14957–14967, 2026.

[49] Richard P. Brent. An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal*, 14(4):422–425, 1971.

[50] Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Conference on Neural Information Processing Systems*, 2023.

[51] Ying-Tu Chen, Wei Hung, Bing-Shu Wu, Zhang-Wei Hong, and Ping-Chun Hsieh. A reward-free viewpoint on multi-objective reinforcement learning. In *International Conference on Learning Representations*, 2026.

[52] Kosuke Nakanishi, Akihiro Kubo, Yuji Yasui, and Shin Ishii. Off-policy actor-critic for adversarial observation robustness: Virtual alternative training via symmetric policy evaluation. In *International Conference on Machine Learning*, pages 45624–45662, 2025.

[53] Roman Belaire, Arunesh Sinha, and Pradeep Varakantham. On minimizing adversarial counterfactual error in adversarial reinforcement learning. In *International Conference on Learning Representations*, 2025.

[54] Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? Towards optimal and efficient evasion attacks in deep RL. In *International Conference on Learning Representations*, 2022.

[55] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[56] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

A Related Work

Our paper is most closely related to three research directions: certified robustness in DRL, zero-shot transfer learning in DRL, and mixture-of-experts (MoE) in DRL. We briefly review these directions and analyze why existing methods are not directly applicable to our setting.

Certified robustness in DRL. Certified robustness in DRL aims to provide formal guarantees for policy behavior under bounded perturbations, rather than only reporting empirical performance under selected attacks. Existing methods [12, 15, 13, 20–27] can be broadly divided into smoothing-based certification, optimization-based certification, and statistical or ensemble-style certification. Smoothing-based methods [12, 15, 13, 20] certify action-level robustness or bound of cumulative reward by applying randomized smoothing to value functions or policies. Optimization-based methods [21–24] derive guarantees through explicit worst-case analysis, such as convex relaxation, interval bound propagation, Lipschitz-constrained policies, or minimax formulations. Other works [25, 20, 15, 26, 27] use hypothesis testing, multi-agent certification, f-divergence programs, or voting-based mechanisms to obtain robustness guarantees. While these methods provide important

certificates for robust DRL, they mainly focus on certifying a single policy within a fixed task. Therefore, when the task changes, their certificates no longer directly apply, making it difficult to preserve certified robustness for the student policy in the target task.

Zero-shot transfer learning in DRL. Zero-shot transfer learning in DRL aims to train policies that can perform well in unseen deployment environments without additional training or data from the test environments [28]. Existing methods can be roughly divided into three categories [28]. The first category reduces the train–test gap by making the training distribution or training objective closer to the expected test conditions, such as through domain randomization [29], environment or task generation [30, 31], or recent augmentation-based visual generalization methods [32, 33]. These methods are usually effective when possible test-time variations are known or can be approximately simulated during training. The second category explicitly models or exploits structural differences between training and testing environments, such as causal or compositional structure and task or goal abstractions [34–36]. These methods usually rely on stronger structural assumptions, including knowing which factors vary across contexts and which factors should remain stable [28]. The third category addresses RL-specific optimization issues, such as improving representation learning, exploration, or policy optimization for better generalization [37, 38]. However, these methods mainly aim to improve the generalization ability of a single policy learned during training. In our setting, the student target preference is unknown during teacher training, and no target-side data or interaction with the target environment is available afterward. Therefore, researchers cannot train or adapt teacher policies through zero-shot generalization to cover the final target preference. As a result, these methods are not directly applicable to our setting.

Mixture-of-experts in DRL. Mixture-of-experts (MoE) architectures provide a natural mechanism for combining multiple expert components through input-dependent gating, expert routing, and weighted aggregation [39]. Recent studies have introduced MoE mechanisms into RL to support modular and scalable policy or value learning across a range of settings [40, 41, 7, 42], including value-based and SoftMoE-style architectures [7, 43], multi-task and visual RL [41, 44, 45], offline and robot-control settings [42, 46], and agentic or vision-language RL [47, 48]. In these studies, MoE components are usually learned or optimized through policy training, value learning, offline trajectories, or routing objectives. However, our setting is training-free: the student aggregates pre-trained teacher policies without target-side data, interaction with the target environment, or additional training. As a result, the student cannot learn a new MoE backbone, optimize expert modules, or train a routing function for the target setting. Moreover, these MoE-based DRL studies do not discuss how certified robustness guarantees can be preserved after aggregating multiple pre-trained certified policies, so they do not directly cover the problem studied in this paper.

B Proofs for Discrete Action Certification

B.1 Restatement and Proof of Thm. 1

Lemma 3 (Lipschitzness of advantage margins)**.** *Assume $|\mathcal{A}| \geq 2$. For teacher i, let $M_{i,1}(s) := \max_{a\in\mathcal{A}} \tilde{Q}_i(s,a)$, and let $M_{i,2}(s)$ denote the second largest value among $\{\tilde{Q}_i(s,a) : a \in \mathcal{A}\}$. Define $m_i(s) := M_{i,1}(s) - M_{i,2}(s)$ and $u_i(s) := \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[m_i(s+\delta)]$. Then both m_i and u_i are $2L_Q$-Lipschitz:*

$$|m_i(s) - m_i(s')| \leq 2L_Q\|s - s'\|_2, \qquad |u_i(s) - u_i(s')| \leq 2L_Q\|s - s'\|_2.$$

Proof. For any states s and s', Lem. 1 gives $\max_{a\in\mathcal{A}} |\tilde{Q}_i(s,a) - \tilde{Q}_i(s',a)| \leq L_Q\|s - s'\|_2$. Since each action value changes by at most $L_Q\|s - s'\|_2$, the largest value and the second largest value among all actions can each change by at most $L_Q\|s - s'\|_2$. Hence both $M_{i,1}$ and $M_{i,2}$ are L_Q-Lipschitz. Since $m_i = M_{i,1} - M_{i,2}$, we obtain $|m_i(s) - m_i(s')| \leq 2L_Q\|s - s'\|_2$. For the smoothed margin, we couple the two expectations with the same Gaussian noise δ. Then $|u_i(s) - u_i(s')| \leq \mathbb{E}_\delta|m_i(s+\delta) - m_i(s'+\delta)| \leq 2L_Q\|s - s'\|_2$. □

Thm. 1 (Discrete certified action bound). *For every fixed action $a \in \mathcal{A}$, the aggregated value $s \mapsto Q_{\text{agg}}(s,a)$ is L_{agg}-Lipschitz with $L_{\text{agg}} := L_Q + \frac{4\beta L_Q}{|\lambda|}$. Let $a^\star_{\text{agg}}(s) \in \arg\max_{a\in\mathcal{A}} Q_{\text{agg}}(s,a)$ be the greedy action of the aggregated student, and define the aggregated action gap as $g_{\text{agg}}(s) := Q_{\text{agg}}(s, a^\star_{\text{agg}}(s)) - \max_{a\in\mathcal{A},\, a\neq a^\star_{\text{agg}}(s)} Q_{\text{agg}}(s,a)$. If $g_{\text{agg}}(s) > 0$, then the student action is certified*

to remain unchanged for every perturbed state s' satisfying $\|s' - s\|_2 < r_{\mathrm{dis}}(s) := \frac{g_{\mathrm{agg}}(s)}{2L_{\mathrm{agg}}}$. That is, $\pi_S(s') = \pi_S(s)$.

Proof. Fix an action $a \in \mathcal{A}$ and write $Q_{\mathrm{agg}}(s,a) = \lambda^{-1} \log \sum_{i=1}^N w_i(s) \exp(\lambda \tilde{Q}_i(s,a))$. Let $z_i(s) := \log w_i(s) + \lambda \tilde{Q}_i(s,a)$. Since the log-sum-exp function is 1-Lipschitz with respect to the ℓ_∞ norm, $|Q_{\mathrm{agg}}(s,a) - Q_{\mathrm{agg}}(s',a)| \le |\lambda|^{-1} \max_i |z_i(s) - z_i(s')|$. The teacher value term is bounded by $|\lambda|\,|\tilde{Q}_i(s,a) - \tilde{Q}_i(s',a)| \le |\lambda| L_Q \|s - s'\|_2$. It remains to control the weight term. Since $w_i(s) = \exp(\beta u_i(s)) / \sum_{k=1}^N \exp(\beta u_k(s))$, we have $\log w_i(s) = \beta u_i(s) - \log \sum_{k=1}^N \exp(\beta u_k(s))$. Using the Lipschitz property of log-sum-exp and Lem. 3, $|\log w_i(s) - \log w_i(s')| \le 4\beta L_Q \|s - s'\|_2$, therefore,

$$|Q_{\mathrm{agg}}(s,a) - Q_{\mathrm{agg}}(s',a)| \le \left(L_Q + \frac{4\beta L_Q}{|\lambda|} \right) \|s - s'\|_2.$$

Thus $Q_{\mathrm{agg}}(\cdot, a)$ is L_{agg}-Lipschitz for every action a. Now let $a^\star_{\mathrm{agg}}(s)$ be the greedy action at s. For any competing action $a \neq a^\star_{\mathrm{agg}}(s)$, the difference $Q_{\mathrm{agg}}(s, a^\star_{\mathrm{agg}}(s)) - Q_{\mathrm{agg}}(s,a)$ can decrease by at most $2L_{\mathrm{agg}}\|s - s'\|_2$ under perturbation. Hence, if $\|s' - s\|_2 < g_{\mathrm{agg}}(s)/(2L_{\mathrm{agg}})$, the greedy action at s still has a larger aggregated value than every competing action at s'. Therefore $\pi_S(s') = \pi_S(s)$. □

C Proofs for Continuous Action Certification

C.1 Restatement and Proof of Lem. 4

Lemma 4 (Bound on peer-review weight)**.** *Let $w_i(s) = \mathbb{P}[f(s+\delta) = i]$, where $\delta \sim \mathcal{N}(0, \sigma^2 I)$. For any s' satisfying $\|s' - s\|_2 \le \epsilon$, define $\alpha_\epsilon := 2\Phi(\epsilon/(2\sigma)) - 1$. Then $\|w(s') - w(s)\|_1 \le 2\alpha_\epsilon$.*

Proof. Let $P_s = \mathcal{N}(s, \sigma^2 I)$ and $P_{s'} = \mathcal{N}(s', \sigma^2 I)$. Then $w(s)$ and $w(s')$ are the pushforward distributions of P_s and $P_{s'}$ under the deterministic map f. By data processing for total variation, $\mathrm{TV}(w(s), w(s')) \le \mathrm{TV}(P_s, P_{s'})$. Let $r = \|s' - s\|_2$. For two Gaussians with common covariance $\sigma^2 I$ and mean distance r, the total variation distance is $2\Phi(r/(2\sigma)) - 1$. This follows by projecting onto the direction $(s' - s)/r$, where the likelihood-ratio test reduces to a one-dimensional Gaussian test with threshold at the midpoint of the two means. Since $r \le \epsilon$, we have $\mathrm{TV}(P_s, P_{s'}) \le 2\Phi(\epsilon/(2\sigma)) - 1 = \alpha_\epsilon$. Finally, for discrete distributions, $\|w(s') - w(s)\|_1 = 2\mathrm{TV}(w(s'), w(s))$. Therefore, $\|w(s') - w(s)\|_1 \le 2\alpha_\epsilon$. □

C.2 Restatement and Proof of Thm. 2

Thm. 2 (Continuous certified action bound). *Fix a state s, perturbation radius ϵ, quantile level $p \in (0,1)$, and action coordinate $k \in [d_a]$. Suppose that, for every teacher $i \in [N]$, Lem. 2 gives certified bounds $L_{i,k}(s,\epsilon)$ and $R_{i,k}(s,\epsilon)$ such that $a_{i,k}(s') \in [L_{i,k}(s,\epsilon), R_{i,k}(s,\epsilon)]$, for all $\|s' - s\|_2 \le \epsilon$. Define $d_{i,k}(s,\epsilon) := R_{i,k}(s,\epsilon) - L_{i,k}(s,\epsilon)$ and*

$$\Delta_k(s,\epsilon) := 2\left(2\Phi\left(\frac{\epsilon}{2\sigma}\right) - 1\right) + \max_{i \in [N]} \left(2\Phi\left(\frac{d_{i,k}(s,\epsilon)}{2\sigma_a}\right) - 1\right).$$

Assume $\Delta_k(s,\epsilon) < \min\{p, 1-p\}$. Let $a_{\mathrm{stu},k,q}(s)$ denote the q-quantile of the clean student mixture CDF $F_{\mathrm{stu},k}(\cdot \mid s)$. Then for every perturbed state s' satisfying $\|s' - s\|_2 \le \epsilon$, $a_{\mathrm{stu},k,p-\Delta_k(s,\epsilon)}(s) \le a_{\mathrm{stu},k,p}(s') \le a_{\mathrm{stu},k,p+\Delta_k(s,\epsilon)}(s)$.

Proof. Fix a coordinate k and write $F_k(\cdot \mid s)$ for $F_{\mathrm{stu},k}(\cdot \mid s)$. For any scalar action value a, add and subtract the mixture with perturbed weights but clean teacher actions:

$$|F_k(a \mid s') - F_k(a \mid s)| \le \sum_{i=1}^N |w_i(s') - w_i(s)| + \sum_{i=1}^N w_i(s') \left| \Phi\left(\frac{a - a_{i,k}(s')}{\sigma_a}\right) - \Phi\left(\frac{a - a_{i,k}(s)}{\sigma_a}\right) \right|.$$

The first term is bounded by $2\alpha_\epsilon$ from Lem. 4. For the second term, both $a_{i,k}(s)$ and $a_{i,k}(s')$ lie in $[L_{i,k}(s,\epsilon), R_{i,k}(s,\epsilon)]$, so their distance is at most $d_{i,k}(s,\epsilon)$. The maximum difference between two Gaussian CDFs with the same variance and centers separated by at most $d_{i,k}(s,\epsilon)$

is $2\Phi(d_{i,k}(s,\epsilon)/(2\sigma_a)) - 1$. Therefore, $|F_k(a \mid s') - F_k(a \mid s)| \le \Delta_k(s,\epsilon)$ for all a. It remains to translate this uniform CDF bound into a quantile bound. Let $\Delta = \Delta_k(s,\epsilon)$. Since $|F_k(a \mid s') - F_k(a \mid s)| \le \Delta$ for all a, any point whose clean CDF value is at least $p + \Delta$ has perturbed CDF value at least p. Hence the perturbed p-quantile is at most the clean $(p+\Delta)$-quantile. Similarly, any point whose clean CDF value is below $p - \Delta$ has perturbed CDF value below p, so the perturbed p-quantile is at least the clean $(p-\Delta)$-quantile. This proves the claim. □

D Proofs for Cumulative Reward Certification

D.1 Restatement and Proof of Thm. 3

Thm. 3 (Vector-valued cumulative reward certificate). *Assume an infinite-horizon discounted MDP with discount factor $\gamma \in [0,1)$. Suppose that the CAST student admits a certified action set $C(s,\epsilon)$. Define $\Delta_A(s,\epsilon) := \sup_{a,b\in C(s,\epsilon)} d_{\mathcal{A}}(a,b)$, $V_{\max}^{\mathrm{vec}} := \frac{R_{\max}^{\mathrm{vec}}}{1-\gamma}$. Under Assump. 1, for any s' satisfying $d_{\mathcal{S}}(s',s) \le \epsilon$,*

$$\|\mathbf{V}^{\pi_S}(s') - \mathbf{V}^{\pi_S}(s)\|_\infty \le \left(L_R^{\mathrm{vec}} + 2\gamma V_{\max}^{\mathrm{vec}} L_P^{\mathrm{TV}}\right)(\epsilon + \Delta_A(s,\epsilon)).$$

Proof. Fix a reward coordinate $\ell \in [m]$ and write $a = \pi_S(s)$ and $a' = \pi_S(s')$. Since $C(s,\epsilon)$ is a certified action set and $d_{\mathcal{S}}(s',s) \le \epsilon$, both a and a' belong to $C(s,\epsilon)$. Hence $d_{\mathcal{A}}(a,a') \le \Delta_A(s,\epsilon)$. By the Bellman equations [9] and the triangle inequality,

$$|V_\ell^{\pi_S}(s) - V_\ell^{\pi_S}(s')| \le |r_\ell(s,a) - r_\ell(s',a')| + \gamma \left|\mathbb{E}_{P(\cdot|s,a)}[V_\ell^{\pi_S}] - \mathbb{E}_{P(\cdot|s',a')}[V_\ell^{\pi_S}]\right|.$$

Assump. 1 gives $|r_\ell(s,a) - r_\ell(s',a')| \le L_R^{\mathrm{vec}} d_{\mathcal{SA}}((s,a),(s',a'))$. Also, $\|V_\ell^{\pi_S}\|_\infty \le V_{\max}^{\mathrm{vec}}$ because the reward is bounded by $R_{\max}^{\mathrm{vec}}$. The total variation inequality for bounded functions then gives

$$\left|\mathbb{E}_{P(\cdot|s,a)}[V_\ell^{\pi_S}] - \mathbb{E}_{P(\cdot|s',a')}[V_\ell^{\pi_S}]\right| \le 2V_{\max}^{\mathrm{vec}} \mathrm{TV}(P(\cdot \mid s,a), P(\cdot \mid s',a')).$$

Using Assump. 1 again, the transition term is bounded by $2V_{\max}^{\mathrm{vec}} L_P^{\mathrm{TV}} d_{\mathcal{SA}}((s,a),(s',a'))$. Therefore,

$$|V_\ell^{\pi_S}(s) - V_\ell^{\pi_S}(s')| \le \left(L_R^{\mathrm{vec}} + 2\gamma V_{\max}^{\mathrm{vec}} L_P^{\mathrm{TV}}\right) d_{\mathcal{SA}}((s,a),(s',a')).$$

Finally, $d_{\mathcal{SA}}((s,a),(s',a')) \le d_{\mathcal{S}}(s,s') + d_{\mathcal{A}}(a,a') \le \epsilon + \Delta_A(s,\epsilon)$. Taking the maximum over reward coordinates proves the theorem. □

Thm. 3 uses the certified action set $C(s,\epsilon)$ to connect action-level certification with cumulative reward certification. For continuous actions, Thm. 2 gives coordinate-wise certified intervals, which define $C_{\mathrm{cont}}(s,\epsilon)$ and the corresponding action diameter $\Delta_A(s,\epsilon)$. For discrete actions, we define $C(s,\epsilon)$ as the set of actions whose clean aggregated values are within $2L_{\mathrm{agg}}\epsilon$ of the clean greedy action value in Lem. 5.

Lemma 5 (Discrete certified action set)**.** *Assume that, for every action $a \in \mathcal{A}$, $Q_{\mathrm{agg}}(\cdot,a)$ is L_{agg}-Lipschitz. For a reference state s and radius $\epsilon \ge 0$, let $a^\star(s) \in \arg\max_{a\in\mathcal{A}} Q_{\mathrm{agg}}(s,a)$ and define*

$$C_{\mathrm{disc}}(s,\epsilon) := \{a \in \mathcal{A} : Q_{\mathrm{agg}}(s,a) \ge Q_{\mathrm{agg}}(s,a^\star(s)) - 2L_{\mathrm{agg}}\epsilon\}.$$

Then $C_{\mathrm{disc}}(s,\epsilon)$ is a certified action set.

Proof. Let s' satisfy $d_{\mathcal{S}}(s',s) \le \epsilon$ and let $a' = \pi_S(s')$. Since a' is greedy at s', $Q_{\mathrm{agg}}(s',a') \ge Q_{\mathrm{agg}}(s',a^\star(s))$. By the Lipschitz bound, $Q_{\mathrm{agg}}(s,a') \ge Q_{\mathrm{agg}}(s',a') - L_{\mathrm{agg}}\epsilon$ and $Q_{\mathrm{agg}}(s',a^\star(s)) \ge Q_{\mathrm{agg}}(s,a^\star(s)) - L_{\mathrm{agg}}\epsilon$. Combining these inequalities gives $Q_{\mathrm{agg}}(s,a') \ge Q_{\mathrm{agg}}(s,a^\star(s)) - 2L_{\mathrm{agg}}\epsilon$. Thus $a' \in C_{\mathrm{disc}}(s,\epsilon)$, so the set is certified. □

D.2 Corollary and Proof

Corollary 1 (Scalarized cumulative reward certificate)**.** *Let $\omega^\star \in \Delta^{m-1}$ be the target preference and define $V_{\omega^\star}^{\pi_S}(s) := (\omega^\star)^\top \mathbf{V}^{\pi_S}(s)$. Suppose the scalarized reward satisfies $|r_{\omega^\star}(s,a) - r_{\omega^\star}(\hat{s},\hat{a})| \le L_R^{\omega^\star} d_{\mathcal{SA}}((s,a),(\hat{s},\hat{a}))$ and $|r_{\omega^\star}(s,a)| \le R_{\max}^{\omega^\star}$ for all (s,a). Let $V_{\max}^{\omega^\star} := \frac{R_{\max}^{\omega^\star}}{1-\gamma}$. Under the transition smoothness condition in Assump. 1 and the certified-action-set condition in Thm. 3, for any s' satisfying $d_{\mathcal{S}}(s',s) \le \epsilon$,*

$$|V_{\omega^\star}^{\pi_S}(s') - V_{\omega^\star}^{\pi_S}(s)| \le \left(L_R^{\omega^\star} + 2\gamma V_{\max}^{\omega^\star} L_P^{\mathrm{TV}}\right)(\epsilon + \Delta_A(s,\epsilon)).$$

Proof. Apply the proof of Thm. 3 to the scalarized reward function $r_{\omega^\star}(s,a) = (\omega^\star)^\top \mathbf{r}(s,a)$. The reward-difference term is controlled by $L_R^{\omega^\star}$, and the value function is bounded by $V_{\max}^{\omega^\star} = R_{\max}^{\omega^\star}/(1-\gamma)$. The transition term is controlled by the same total-variation smoothness constant L_P^{TV}. Combining these terms gives the stated bound. □

E Certification Backbone and Baseline Choices

We clarify our choices of certification backbones for constructing the certified teacher policies and empirical baselines for evaluating CAST. For discrete-action tasks, we build on CROP [12], a classic smoothing-based certified DRL method for discrete policies that smooths action-value functions and provides both action-level and cumulative-reward certificates. Although [13] provides a more recent smoothing technique for discrete policies, its technique mainly preserves the information of the best action while suppressing the information of other actions. This design is not suitable for our adaptation setting, because an action that is suboptimal under a source reward preference may become useful under a changed target reward preference. Therefore, CAST retains the action-value information of all candidate actions and uses it for robustness-aware aggregation and certification. For continuous-control tasks, we only borrow the quantile-smoothing certification technique from [13], a recent state-of-the-art smoothing-based certified DRL work that provides both action-level and cumulative-reward-level robustness bounds for continuous-control tasks. We do not use its other components because one of our goals is to isolate whether certified robustness can be transferred from smoothed teachers to the CAST student. Introducing additional components would create confounding factors and make the transfer effect harder to analyze. The above methods are used as certification backbones for the teachers and for evaluating the smoothed student.

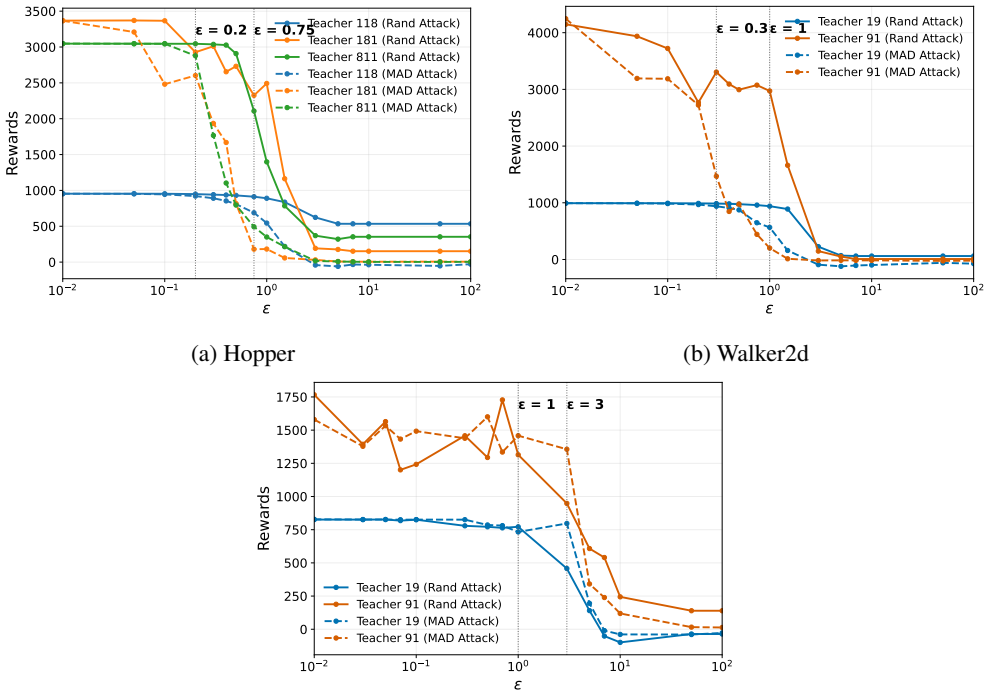
We next explain our choice of adaptation baselines. Based on the related work, MoE-based adaptation and zero-shot transfer are the two method families most closely related to our setting. We do not include MoE-based adaptation methods as baselines because they usually require target-side data, environment interaction, or additional training to learn routing and adaptation, while CAST directly aggregates pre-trained teacher policies without additional training. Moreover, existing MoE-based DRL methods do not provide certified robustness guarantees, so they are not directly comparable under our certified robustness transfer setting. The closest available comparison to our no-retraining setting is zero-shot policy adaptation, which produces a policy for a specified target preference without further training at deployment. We therefore include MORL-FB [51] as a zero-shot empirical baseline. Nevertheless, MORL-FB does not exactly match our setting for two reasons: (1) it requires interaction with the target environment to collect target-side data for training, whereas CAST starts from fixed certified source teachers without any target-side data or environment interaction; and (2) it does not provide certified robustness guarantees for the resulting target policy.

F Experimental Parameter Selection

Parameter Selection. We select the evaluation, attack, and certification parameters through preliminary parameter scan before reporting the main results. This procedure is intended to reflect a practical use case of CAST, where the parameters should be chosen by balancing clean utility, attack strength, certification quality, and computational cost. The selected quantities include perturbation budgets ϵ for each environment, smoothing scales σ for smoothed teacher policies, Monte Carlo smoothing sample size m for smoothed evaluation and certification, action-space smoothing scale σ_a for the mixture-based continuous student, SARSA attacker regularization coefficient `reg`, and discrete aggregation parameters β and λ for the advantage-margin weighted log-sum-exp rule.

F.1 Continuous-Control Parameter Selection

Perturbation budgets. Different continuous-control environments have different sensitivities to state perturbations, so we choose different perturbation budgets for different environments rather than using a single ϵ for all tasks. Fig. 3 shows the performance of clean policies under Rand and MAD attacks as the perturbation radius increases. Based on this preliminary sweep, we select one weaker and one stronger perturbation budget for each environment: $\epsilon = 0.2$ and 0.75 for Hopper, $\epsilon = 0.3$ and 1.0 for Walker2d, and $\epsilon = 1.0$ and 3.0 for Humanoid.



(a) Hopper

(b) Walker2d

(c) Humanoid

Figure 3: Performance of trained teacher policies under Rand and MAD attacks with increasing state perturbation radius ϵ on continuous-control tasks.

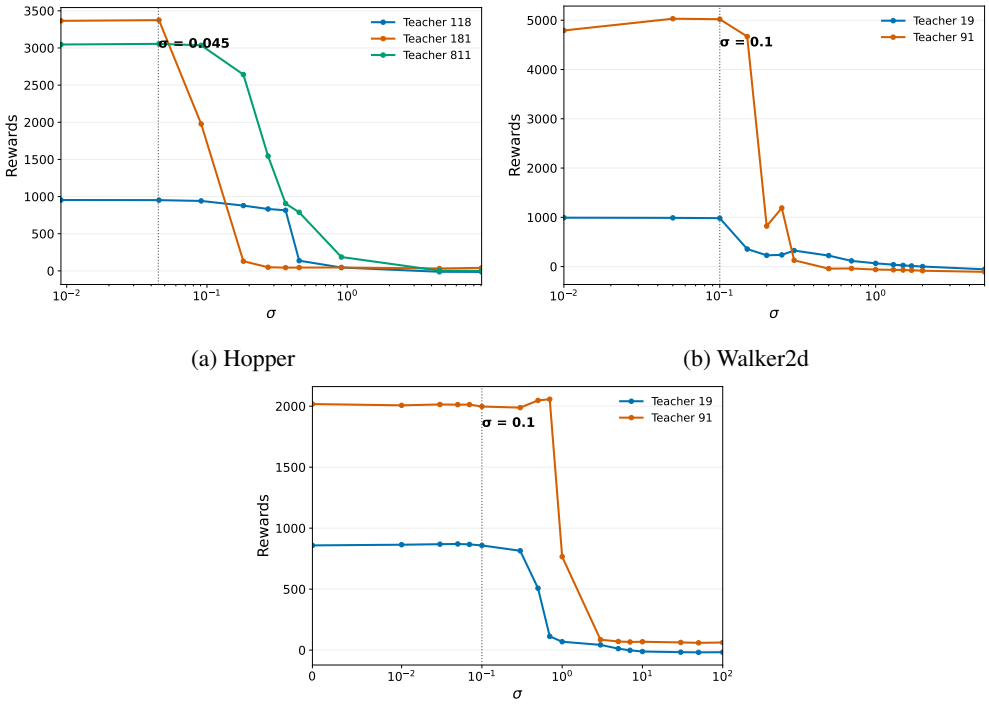
State smoothing scale selection and validation under teacher and student preferences. We first select the state smoothing scale σ based on the sensitivity of the smoothed teacher policies under their source preferences. We then validate the selected scales on the aggregated student under the target preference and different perturbation magnitudes.

For the teacher policies, we select the largest σ that does not degrade clean performance, since a larger smoothing scale can improve robustness but may reduce clean utility. Fig. 4 reports the clean performance under different values of σ. Based on this sweep, we select $\sigma = 0.045$ for Hopper and $\sigma = 0.1$ for Walker2d and Humanoid.

For the aggregated student, Tab. 4 evaluates the selected scales under the student target preference and different perturbations. These results show that the scales selected from the teacher-side analysis also maintain stable student performance. Specifically, on Hopper, the selected $\sigma = 0.045$ lies within the transition region before performance degrades; on Walker2d, performance becomes more stable for $\sigma \geq 0.07$, supporting the choice of $\sigma = 0.1$; and on Humanoid, increasing σ provides no consistent improvement, while the selected $\sigma = 0.1$ lies between the evaluated values 0.07 and 0.13, which exhibit similar trends.

Number of smoothing samples. Following [13], we evaluate different numbers of smoothing samples m used in Monte Carlo estimation. Although a very large sample size such as $m = 10000$ can reduce estimation variance, it is not a practical setting for implementation [13]. Tab. 5 shows that $m = 100$ gives stable rewards compared with larger sample sizes. We therefore use $m = 100$ for continuous control evaluations.

Action-space smoothing scale. For the mixture-based continuous student, we further tune the action smoothing scale σ_a. Fig. 5 reports student performance under different σ_a values, together with the cumulative reward averaged over the source-teacher preferences. We select the value that best preserves reward under both the aggregated student and source-teacher preferences. The student-preference reward and the average source-preference reward exhibit consistent trends. In particular, on Hopper, they peak around $\sigma_a = 10^{-4}$ and decrease as σ_a increases; on Walker2d, they increase



(a) Hopper

(b) Walker2d

(c) Humanoid

Figure 4: Performance of smoothed policies without attacks for different state smoothing scales σ on continuous control tasks.

with σ_a and reach their highest values around $\sigma_a = 1$; and on Humanoid, they peak at $\sigma_a = 10^{-4}$. Based on these results, we set $\sigma_a = 10^{-4}$ for Hopper and Humanoid, and $\sigma_a = 1$ for Walker2d.

SARSA attacker regularization. For SARSA attacks, we tune the attacker regularization coefficient `reg`. In our implementation, `reg` is the weight of the regularization term in the SARSA attacker training loss:

$$\texttt{total_loss} = \texttt{td_loss} + \texttt{reg} \cdot \texttt{reg_loss}.$$

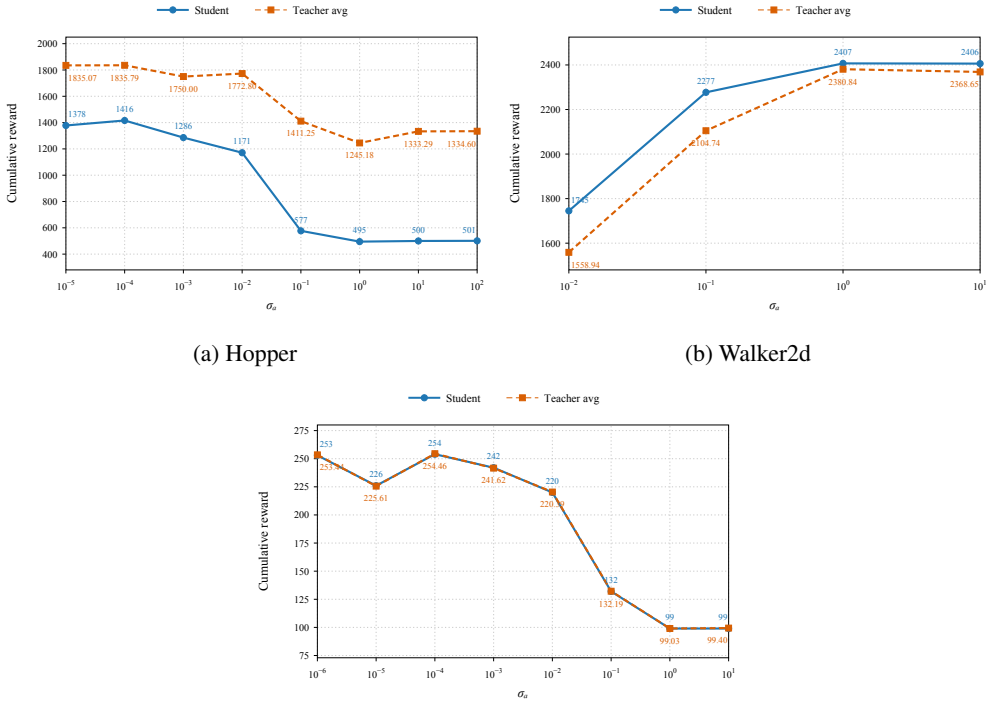
Thus, `reg` controls the tradeoff between the SARSA temporal difference loss and the regularization term when training the attacker. We scan `reg` values from {0.1, 0.3, 1.0, 3.0, 10.0}. We choose `reg` by jointly considering its attack effect on each teacher. This criterion avoids favoring a setting that is very strong for some teachers but weak for others. Tab. 6 reports this scan across continuous control teachers and perturbation budgets. Based on these results, we use `reg`=3.0 for Hopper and Walker2d, and `reg`=0.1 for Humanoid in the main attack evaluations.

F.2 Discrete-Control Parameter Selection

Perturbation budgets. For discrete environments, we also use environment specific perturbation budgets. Fig. 6 shows how policy performance changes as ϵ increases under Rand and FGSM attacks. Based on this trend, we select one weaker budget and one stronger budget for each environment. Specifically, we set $\epsilon \in \{0.7, 3.0\}$ for Deep Sea and $\epsilon \in \{0.3, 1.0\}$ for Fruit Tree.

State smoothing scale selection and validation under teacher and student preferences. We first select the state smoothing scale σ based on the sensitivity of the smoothed teacher policies under their source preferences. We then validate the selected scale on the aggregated student under the target preference and different perturbations.

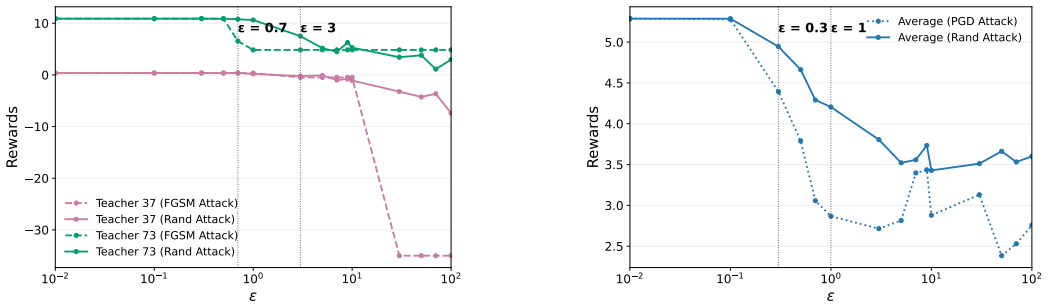
For the teacher policies, Fig. 7 shows that increasing σ can improve certification strength but may reduce policy utility. Based on this tradeoff, we select $\sigma = 0.5$ for the discrete-control evaluations.



(a) Hopper (b) Walker2d

(c) Humanoid

Figure 5: Student-preference and teacher-preference average cumulative rewards under different action smoothing scales σ_a.



(a) Deep Sea Treasure (b) Fruit-Tree

Figure 6: Performance of trained teacher policies under Rand and FGSM attacks with increasing state perturbation budgets ϵ on discrete control tasks (for Fruit Tree, we report the average teacher performance because the number of teachers is large).

Table 4: Student reward sensitivity to the state-space smoothing scale σ and perturbation radius ϵ on Hopper, Walker2d, and Humanoid under Rand and MAD attacks.

Env.	Attack	σ	Perturbation radius			
			ϵ_1	ϵ_2	ϵ_3	ϵ_4
$\epsilon_1 = 0.01,\ \epsilon_2 = 0.05,\ \epsilon_3 = 0.09,\ \epsilon_4 = 0.13$ for Hopper						
Hopper	Rand	0.005	1835.75 ± 34.65	1895.45 ± 46.27	1729.72 ± 119.30	1617.87 ± 56.48
		0.030	1751.05 ± 49.14	1895.56 ± 291.77	1694.82 ± 192.00	1794.03 ± 167.60
		0.060	1672.01 ± 306.61	1601.36 ± 235.55	1471.38 ± 195.04	1660.76 ± 279.57
		0.090	1544.55 ± 328.64	1507.57 ± 300.59	1468.42 ± 251.19	1684.43 ± 266.27
	MAD	0.005	1776.40 ± 30.11	1751.90 ± 187.35	1744.91 ± 199.59	1770.96 ± 121.55
		0.030	1811.60 ± 184.73	1798.84 ± 100.97	1864.74 ± 83.65	1705.51 ± 222.71
		0.060	1462.84 ± 200.44	1803.72 ± 191.34	1890.20 ± 84.63	1790.93 ± 99.67
		0.090	1725.04 ± 273.80	1785.60 ± 190.29	1705.55 ± 278.11	1682.47 ± 207.15
$\epsilon_1 = 0.01,\ \epsilon_2 = 0.05,\ \epsilon_3 = 0.09,\ \epsilon_4 = 0.13$ for Walker2d						
Walker2d	Rand	0.01	982.04 ± 584.94	678.41 ± 598.98	964.65 ± 611.54	845.26 ± 486.33
		0.07	1345.92 ± 521.38	1577.74 ± 428.56	1481.32 ± 427.43	1505.10 ± 494.52
		0.13	1508.56 ± 527.76	1747.01 ± 282.58	1773.11 ± 327.80	1753.84 ± 240.57
		0.19	1983.95 ± 330.19	1910.59 ± 51.27	1816.31 ± 279.61	1894.84 ± 29.50
	MAD	0.01	2322.98 ± 5.37	2318.17 ± 19.16	1702.34 ± 826.54	1671.96 ± 782.85
		0.07	2341.21 ± 8.45	2340.18 ± 15.82	2326.25 ± 7.07	2276.83 ± 5.63
		0.13	2377.31 ± 10.86	2374.73 ± 14.93	2318.65 ± 24.12	2290.37 ± 17.95
		0.19	2400.17 ± 9.02	2390.15 ± 10.52	2356.24 ± 17.15	2307.67 ± 10.34
$\epsilon_1 = 0.4,\ \epsilon_2 = 0.7,\ \epsilon_3 = 1.0,\ \epsilon_4 = 1.3$ for Humanoid						
Humanoid	Rand	0.01	78.14 ± 15.98	60.33 ± 11.25	49.86 ± 8.58	42.89 ± 5.73
		0.07	71.85 ± 14.13	61.09 ± 14.19	47.04 ± 5.89	37.09 ± 9.39
		0.13	89.47 ± 30.28	56.49 ± 9.67	41.25 ± 10.91	39.70 ± 7.05
		0.19	91.97 ± 21.26	56.24 ± 12.31	42.69 ± 14.57	36.84 ± 8.40
	MAD	0.01	25.70 ± 7.52	21.46 ± 11.34	11.65 ± 6.90	11.32 ± 11.71
		0.07	23.77 ± 8.37	17.35 ± 8.96	11.22 ± 10.69	10.22 ± 7.85
		0.13	27.25 ± 8.71	13.06 ± 8.48	8.11 ± 8.17	5.22 ± 11.47
		0.19	22.58 ± 10.85	12.16 ± 10.54	8.26 ± 4.81	11.67 ± 9.82

For the aggregated student, Tab. 7 shows that, on Deep Sea Treasure, $\sigma = 0.5$ maintains stable rewards under both Rand and PGD across all evaluated values of ϵ, whereas increasing the scale to $\sigma = 1.0$ leads to reward degradation. On Fruit-Tree, $\sigma = 0.5$ also maintains stable student performance across perturbation magnitudes and attacks. These results show that the scale selected from the teacher-side analysis also preserves stable student utility under the target preference and state perturbations.

Number of smoothing samples. Tab. 8 evaluates different numbers of smoothing samples for discrete teachers. Similar to the continuous setting, $m = 100$ provides a good balance between Monte Carlo estimation reliability and computational cost. We therefore use $m = 100$ for discrete control evaluations.

Discrete aggregation parameters. We tune the discrete aggregation parameters β and λ. Tab. 9 shows that $\beta = 1$ and $\lambda = 1$ achieve the best aggregation reward on Fruit Tree while maintaining the same reward on Deep Sea Treasure. We therefore use this setting in the remaining discrete experiments.

Table 5: Effect of Monte Carlo smoothing sample size m on continuous-control teacher evaluation.

Environment	Teacher	$m = 10$	$m = 100$	$m = 1000$	$m = 10000$
Hopper	118	943.12 ± 0.30	952.13 ± 0.06	953.05 ± 0.01	953.14 ± 0.01
	181	3358.51 ± 14.00	3374.29 ± 14.63	3386.22 ± 2.64	3386.89 ± 3.46
	811	2785.63 ± 770.85	3055.18 ± 4.82	3056.21 ± 2.68	3056.95 ± 2.73
	mean	2362.42	2460.54	2465.16	2465.66
Walker2d	19	969.31 ± 1.06	982.92 ± 0.28	984.06 ± 0.17	984.44 ± 0.010
	91	4415.90 ± 722.38	5020.85 ± 29.71	5013.99 ± 60.37	5033.54 ± 51.63
	mean	2692.61	3001.89	2999.03	3008.99
Humanoid	19	809.61 ± 80.23	858.37 ± 3.94	862.41 ± 3.51	861.58 ± 3.04
	91	1999.65 ± 8.37	1997.28 ± 26.29	1998.03 ± 28.33	2023.45 ± 25.91
	mean	1404.63	1427.82	1430.22	1442.52

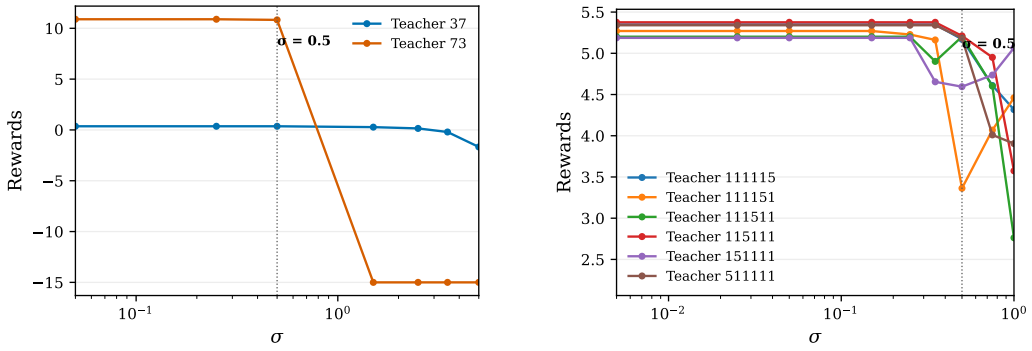
Table 6: SARSA attack rewards for continuous-control teachers under different attacker regularization coefficients (`reg`).

Environment	ϵ	Model	`reg`=0.1	`reg`=0.3	`reg`=1.0	`reg`=3.0	`reg`=10.0
Hopper	0.2	teacher-118	851.24	853.12	853.60	845.77	849.87
		teacher-181	2208.33	2209.09	2197.51	2191.54	2106.24
		teacher-811	2276.45	2223.84	1913.45	2240.51	2071.99
	0.75	teacher-118	807.11	760.32	769.13	748.47	768.91
		teacher-181	1508.59	1016.35	248.62	158.93	421.35
		teacher-811	555.66	207.81	230.07	260.73	337.81
Walker2d	0.3	teacher-91	554.91	621.11	763.22	842.02	768.78
		teacher-19	991.90	988.77	989.77	989.30	988.87
	1.0	teacher-91	-23.17	-26.19	2.83	-52.17	6.68
		teacher-19	991.29	990.52	916.40	499.68	992.88
Humanoid	1.0	teacher-19	2.93	0.75	1.71	0.12	3.51
		teacher-91	-7.01	-3.18	-1.83	-1.70	-4.50
	3.0	teacher-19	-1.27	2.43	-1.75	-0.33	-1.65
		teacher-91	-4.65	-3.75	-5.08	-4.89	-6.53

G Additional Experimental Details

This appendix collects implementation details that support reproducibility. It includes runtime, computational resources, and training details for continuous-action and discrete-action experiments.

Estimating $L_R^{\omega^\star}$ and L_P^{TV}. Since policy performance is evaluated using the scalarized cumulative reward under the target preference $\omega^\star$, we use the scalarized bound in Cor. 1, which specializes the vector-valued certificate in Thm. 3 to this evaluation objective. Let $x_i = (s_i, a_i)$ and $d_{\mathcal{SA}}(x_i, x_j) = \|s_i - s_j\|_2 + \|(a_i - a_j) \oslash (a_{\max} - a_{\min})\|_2$. To instantiate this bound empirically, we estimate the required constants from a local transition dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i^+)\}$ collected around trajectories of the evaluated policies. We estimate $\hat{L}_R^{\omega^\star}$ as the 0.99 quantile of $|(\omega^\star)^\top (r_i - r_j)|/(d_{\mathcal{SA}}(x_i, x_j) + \tau)$, where $\tau > 0$ is a small stabilizer, and set $\hat{R}_{\max}^{\omega^\star} = \max_i |(\omega^\star)^\top r_i|$ and $\hat{V}_{\max}^{\omega^\star} = \hat{R}_{\max}^{\omega^\star}/(1-\gamma)$. For transitions, directly estimating L_P^{TV} is unstable in nearly deterministic MuJoCo dynamics, so we use a Gaussian surrogate $\hat{P}_{\theta,\sigma_{\mathrm{dyn}}}(\cdot \mid s, a) = \mathcal{N}(\mu_\theta(s,a), \sigma_{\mathrm{dyn}}^2 I)$. We estimate the local mean-dynamics Lipschitz constant $\hat{L}_\mu$ as the 0.99 quantile of $\|\mu_\theta(x_i) - \mu_\theta(x_j)\|_2/(d_{\mathcal{SA}}(x_i, x_j) + \tau)$, and set $\hat{L}_P^{\mathrm{TV}} = \hat{L}_\mu/(2\sigma_{\mathrm{dyn}})$ using the Gaussian TV bound, with a fixed surrogate dynamics noise level $\sigma_{\mathrm{dyn}} = 0.3$ across all MuJoCo tasks to stabilize TV estimation under nearly deterministic dynamics.



(a) Deep Sea Treasure

(b) Fruit-Tree

Figure 7: Smoothed-policy performance under different state smoothing scales σ on discrete-control tasks.

Table 7: Student-level reward sensitivity to the state-space smoothing scale σ and perturbation radius ϵ in discrete-action environments under Rand and PGD attacks.

Environment	Attack	σ	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.7$
Deep Sea Treasure	Rand	0.05	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00	2.53 ± 0.35
		0.10	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00	2.32 ± 0.35
		0.50	2.35 ± 0.00	2.46 ± 0.16	2.39 ± 0.10	2.46 ± 0.16
		1.00	2.39 ± 0.10	-0.28 ± 8.24	2.39 ± 0.10	-0.28 ± 8.24
	PGD	0.05	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00	1.35 ± 0.00
		0.10	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00	1.91 ± 0.86
		0.50	2.42 ± 0.14	2.39 ± 0.10	2.39 ± 0.10	2.35 ± 0.00
		1.00	2.39 ± 0.11	-3.12 ± 10.94	-3.12 ± 10.94	-8.48 ± 13.50
Fruit-Tree	Rand	0.05	3.39 ± 0.08	3.35 ± 0.22	3.23 ± 0.29	3.21 ± 0.32
		0.10	3.50 ± 0.06	3.49 ± 0.06	3.49 ± 0.06	3.41 ± 0.21
		0.50	3.59 ± 0.05	3.55 ± 0.13	3.56 ± 0.14	3.49 ± 0.21
		1.00	3.69 ± 0.08	3.70 ± 0.07	3.69 ± 0.09	3.62 ± 0.10
	PGD	0.05	3.53 ± 0.03	3.60 ± 0.05	3.85 ± 0.00	3.20 ± 0.50
		0.10	3.42 ± 0.09	3.63 ± 0.02	3.85 ± 0.00	3.41 ± 0.61
		0.50	3.60 ± 0.08	3.40 ± 0.27	3.48 ± 0.28	3.50 ± 0.31
		1.00	3.68 ± 0.09	3.65 ± 0.11	3.57 ± 0.09	3.53 ± 0.16

G.1 Runtime and Computational Resources

All experiments were conducted on a Linux workstation equipped with an NVIDIA GeForce RTX 5090 GPU with 32GB memory. The workstation used Ubuntu 24.04.4 LTS with Linux kernel 6.17.0-20-generic, NVIDIA driver 580.105.08, and CUDA 13.0.

Tab. 10 evaluates runtime from three perspectives. The full-episode results compare the end-to-end wall-clock costs of an individual teacher, uniform-mean aggregation, and CAST, although these values are also affected by the actual episode lengths of the policies. Taken together, the full-episode and component timings show that the additional CAST runtime in continuous-action environments mainly arises from student aggregation, which requires numerically solving for the quantiles of the weighted teacher mixture. In discrete-action environments, the main computational cost instead comes from average teacher fetching, while the student aggregation overhead remains small. The fixed-state results further show that runtime increases only moderately with the smoothing sample size m in continuous-action environments but grows in discrete-action environments.

Table 8: Effect of Monte Carlo smoothing sample size m on discrete-control teacher evaluation.

Environment	Teacher	$m = 10$	$m = 100$	$m = 1000$	$m = 10000$
Deep Sea Treasure	37	0.36 ± 0.00	0.36 ± 0.00	0.36 ± 0.00	0.36 ± 0.00
	73	10.77 ± 0.15	10.83 ± 0.12	10.89 ± 0.00	10.89 ± 0.00
	mean	5.56	5.59	5.63	5.63
Fruit-Tree	511111	4.61 ± 0.73	5.18 ± 0.47	5.34 ± 0.00	5.34 ± 0.00
	151111	4.32 ± 0.68	4.60 ± 0.00	4.60 ± 0.00	4.60 ± 0.00
	115111	4.85 ± 0.78	5.21 ± 0.20	5.19 ± 0.19	5.06 ± 0.16
	111511	4.00 ± 1.47	5.20 ± 0.00	5.20 ± 0.00	5.20 ± 0.00
	111151	4.89 ± 0.51	3.36 ± 1.53	2.89 ± 1.49	1.91 ± 0.00
	111115	4.25 ± 1.11	5.17 ± 0.23	5.35 ± 0.00	5.35 ± 0.00
	mean	4.49	4.79	4.76	4.58

Table 9: Clean student aggregation under different discrete aggregation parameters (β, λ) with $\sigma = 0.5$ and $m = 100$. Each entry reports the student-preference cumulative reward followed by the source-preference average cumulative reward, separated by "/".

Deep Sea Treasure				
$\beta \backslash \lambda$	0.01	0.1	1	10
0.01	2.3850/2.3850	2.3850/2.3850	2.3850/2.3850	2.3850/2.3850
0.1	2.3850/2.3850	2.3850/2.3850	2.3850/2.3850	2.3850/2.3850
1	2.3850/2.3850	2.3850/2.3850	**2.3850/2.3850**	2.3850/2.3850
10	2.3850/2.3850	2.3850/2.3850	2.3850/2.3850	2.3850/2.3850
Fruit-Tree				
$\beta \backslash \lambda$	0.01	0.1	1	10
0.01	3.5415/3.4863	3.5415/3.5415	3.5631/3.5631	3.5598/3.5598
0.1	3.4863/3.4863	3.4863/3.4863	3.5631/3.5631	3.5598/3.5598
1	3.4092/3.4092	3.4978/3.4978	**3.5936/3.5936**	3.5373/3.5373
10	3.4339/3.4339	3.4339/3.4339	3.4782/3.4782	3.4739/3.4739

G.2 Training Details for Continuous-Action DRL

For the continuous-control experiments, all designated teachers are trained with soft actor-critic (SAC). The teacher policy is a Gaussian MLP with two hidden layers of width 256 and ReLU activations, followed by a tanh squashing function. The critic uses twin Q-networks, each with the same 256-256 ReLU MLP structure, and training uses the minimum of the two critics for target computation. State normalization is enabled for all environments, reward normalization is disabled, and prioritized replay is used throughout training. The details are shown in Tab. 11.

For adversarial training/evaluation, we use two learned attack models. The SARSA attacker is a scalar $Q(s, a)$ regressor, while PAAD is a PPO-style adversarial policy-value pair. For student attacks, the attacker architecture is unchanged; only the scalarization target switches from the designated teacher preference to the student preference. In the main experiments, the selected SARSA regularization coefficient is 3.0 for Hopper and Walker2d, and 0.1 for Humanoid. The details are shown in Tab. 12.

G.3 Training Details for Discrete-Action DRL

For the discrete environments, all designated teachers are trained with double DQN using a fully connected Q-network. The network takes the flattened state as input and uses three hidden layers with ReLU activations. A target network is maintained with soft updates. The details are shown in Fig. 13.

For smoothed discrete attacks, the victim policy is the smoothed teacher or the smoothed aggregated student, both built with training-time smoothing parameter $m = 1$ and test-time Gaussian smoothing centered at $\sigma = 0.5$. Student aggregation uses $\beta = 1.0$ and $\lambda = 1.0$ in both Deep Sea Treasure and

Table 10: Runtime evaluation across environments.

Full-episode runtime (ms), $m = 100$			
Environment	Individual Teacher	Uniform Mean	CAST
Hopper	309.34 ± 18.26	630.10 ± 25.20	4234.76 ± 41.83
Walker2d	312.20 ± 20.69	439.92 ± 124.08	3766.17 ± 19.87
Humanoid	699.14 ± 6.91	85.27 ± 7.62	7032.50 ± 2274.25
Deep Sea Treasure	3.00 ± 2.17	10.02 ± 0.51	414.93 ± 19.66
Fruit-Tree	1.63 ± 0.12	10.89 ± 0.71	425.44 ± 12.08
CAST component timing (ms/step)			
Environment	Average teacher fetching	Student aggregation	Component sum
Hopper	0.45 ± 0.16	3.90 ± 1.14	4.36 ± 1.21
Walker2d	0.63 ± 0.22	3.72 ± 1.02	4.35 ± 1.16
Humanoid	8.65 ± 2.05	23.56 ± 6.40	32.21 ± 8.04
Deep Sea Treasure	14.73 ± 2.56	0.10 ± 0.20	14.84 ± 2.60
Fruit-Tree	14.58 ± 0.55	0.10 ± 0.02	14.68 ± 0.56
Fixed-state CAST runtime (ms/step)			
Environment	$m = 10$	$m = 100$	$m = 1000$
Hopper	3.91 ± 0.13	3.98 ± 0.21	6.83 ± 0.65
Walker2d	2.99 ± 0.29	3.52 ± 0.12	5.23 ± 0.53
Humanoid	11.28 ± 1.49	16.93 ± 1.36	24.15 ± 0.86
Deep Sea Treasure	0.83 ± 0.09	27.07 ± 0.58	2403.91 ± 31.20
Fruit-Tree	2.89 ± 0.33	92.19 ± 16.14	8162.51 ± 54.06

Fruit-Tree. The learned PAAD attacker is a Gaussian director network with a shared backbone and separate policy/value heads. The details are shown in Tab. 14.

H Full Student Preference Utility under State Perturbations

Continuous-control tasks. Tab. 15 compares the smoothed CAST student with the certified teachers and MORL-EB under state perturbations. Across the twenty-four attack settings, CAST exceeds the best teacher in thirteen settings and performs at or above the lowest-reward teacher in twenty-three settings. CAST also outperforms MORL-FB in twenty-one of the twenty-four configurations. Moreover, despite using target-side data and environment interaction during training, MORL-FB falls below the lowest-reward teacher in fifteen of the twenty-four configurations, whereas CAST does so in only one. These results suggest that CAST can reliably transfer useful teacher behavior to the target preference, largely avoid negative transfer, and reduce dependence on any single teacher policy without target-side information. Some student results show relatively large standard deviations. We observe that these cases usually also have large standard deviations for the corresponding teachers. This suggests that the student variance is largely inherited from the teacher variance, rather than caused by instability of the aggregation rule itself. The teacher variance arises because the source teachers are trained with different preferences and are directly evaluated under the student target preference without target side retraining. This mismatch can lead to larger trajectory level fluctuations across repeated evaluations. State attacks further amplify this effect by pushing policies into perturbed regions.

Discrete control tasks. Tab. 16 compares the smoothed CAST student with the source teachers and MORL-FB on Deep Sea Treasure and Fruit Tree. Across the sixteen attack configurations, CAST performs within the teacher reward range in twelve configurations and exceeds the best teacher in the remaining four. Specifically, CAST outperforms both teachers on Deep Sea Treasure

Table 11: Training configuration of the original continuous-action SAC teachers (all designated teachers within the same environment share the same network architecture and optimization hyperparameters).

Item	Setting
Teacher policy network	Gaussian SAC policy, MLP with hidden sizes $(256, 256)$, ReLU activations, tanh-squashed output
Teacher critic network	Twin Q-networks, each an MLP with hidden sizes $(256, 256)$ and ReLU activations
Replay buffer	Prioritized replay; capacity 10^6; warm-up size 10^4; priority exponent $\alpha = 0.5$; initial importance-sampling coefficient $\beta = 0.4$
Optimization	Adam optimizer; batch size 256; discount factor $\gamma = 0.99$; one gradient update per environment step
Target update	Soft target update every step with mixing rate $\tau = 0.005$
Entropy regularization	Soft-Q training with automatic temperature tuning; target entropy is 0.2 for Hopper and "dim" for Walker2d/Humanoid
Learning rate	Hopper/Walker2d: 3×10^{-4} for critic, actor, and temperature; Humanoid: 1×10^{-4} for critic, actor, and temperature
Training horizon	5×10^5 frames for Hopper, 1.5×10^6 frames for Walker2d, and 3×10^6 frames for Humanoid

Table 12: Training configuration of the learned continuous-action attackers (the PAAD architecture is identical for teacher-targeted and student-targeted attacks, and only the attack objective changes).

Attacker	Network structure	Training hyperparameters
SARSA	MLP $Q(s, a)$ with hidden sizes $(256, 256)$, ReLU activations, scalar output	30 SARSA steps, rollout horizon 2048, 10 value epochs per step, 32 minibatches, learning rate 10^{-3}, $\gamma = 0.99$, Gaussian exploration std. 0.1, action-change ratio 0.5, and ℓ_2 perturbation constraint
PAAD policy	Gaussian PPO adversary with hidden sizes $(64, 64)$, tanh activations, orthogonal initialization	1220 training steps, PPO epochs $= 3$, 32 minibatches, policy learning rate 10^{-2}, clip ratio 0.4, entropy coefficient 3×10^{-4}, GAE parameter $\lambda = 0.95$, and one inner attack step
PAAD value	Value MLP with hidden sizes $(64, 64)$, tanh activations, orthogonal initialization	Value learning rate 10^{-3}, discount factor $\gamma = 0.99$, rollout horizon 512 for Hopper and 521 for Walker2d/Humanoid

under Rand, FGSM, and PGD at $\epsilon = 3.0$, and on Fruit Tree under PAAD at $\epsilon = 0.3$. CAST also outperforms MORL-FB in ten configurations and ties it in one. Despite using target-side data and environment interaction during training, MORL-FB performs below the lowest-reward teacher in eight configurations and ties the lowest-reward teacher in one, whereas CAST never falls below the lowest-reward teacher. These results show that CAST can reliably recover and preserve useful target-preference behavior from the source teachers while largely avoiding negative transfer without target-side information.

I Additional Target-Preference Sensitivity Experiments

This appendix reports the full target-preference sensitivity study. The goal is to test whether CAST preserves target-preference utility when the target preference changes beyond the balanced setting used in the main benchmark. The tables below use the same teacher and student notation as Sec. 6, and each entry reports the student-preference scalarized return mean and standard deviation under representative attacks.

Target-Preference Sensitivity. We further evaluate CAST under multiple student preferences across all five environments to test target-preference sensitivity beyond the balanced target preference. Since this study is not intended to repeat the full attack benchmark, we report representative attacks and perturbation levels for each environment. As shown in Tabs. 17, 18, 19, 20, and 21, CAST stays within the teacher performance range in $57/116$ cases and exceeds the best teacher in $44/116$ cases. Therefore, in $101/116$ cases, CAST either remains within the teacher envelope or improves beyond

Table 13: Training configuration of the original discrete-action DQN teachers (all designated teachers within the same environment share the same architecture and optimization hyperparameters).

Item	Setting
Teacher network	Fully connected DQN with hidden sizes $(512, 512, 512)$ and ReLU activations
Target network	One target Q-network with soft update
Replay buffer	Uniform replay buffer with capacity 10^6 transitions
Optimization	Adam optimizer; batch size 256; learning rate 3×10^{-4}
Bellman update	Discount factor $\gamma = 0.995$; target-update rate $\tau = 0.005$; one update every 5 environment steps
Exploration	Uniformly random actions for the first 10^4 steps, then fixed ϵ-greedy exploration with $\epsilon = 0.35$
Evaluation	Evaluation every 10^4 steps on 5 episodes; checkpoints saved every 5×10^4 steps
Training horizon	The released teacher checkpoints for both Deep Sea Treasure and Fruit-Tree use 1,005,000 training steps

Table 14: Training configuration of the learned PAAD attacker in the discrete-control experiments (there is no separately trained SARSA attacker in the discrete pipeline).

Component	Network structure	Training hyperparameters
PAAD director	MLP backbone with two tanh hidden layers; a Gaussian mean head over a $(\lvert\mathcal{A}\rvert - 1)$-dimensional raw action, a learned log-standard-deviation vector, and a scalar value head	PPO-style training with $\gamma = 0.99$, GAE $\lambda = 0.95$, clip ratio 0.4, entropy coefficient 3×10^{-4}, value-loss coefficient 0.5, max gradient norm 0.5, and learning rate 3×10^{-4}
Deep Sea PAAD	Hidden size 128 for both hidden layers	2441 training steps, rollout horizon 2048, 3 PPO epochs, 32 minibatches, 1 inner attack step, and training attack radius $\epsilon \in \{0.7, 3.0\}$
Fruit PAAD	Hidden size 512 for both hidden layers	1220 training steps, rollout horizon 2048, 3 PPO epochs, 32 minibatches, 5 inner attack steps, and training attack radius $\epsilon \in \{0.3, 1.0\}$
Teacher/student target	Same PAAD architecture for teacher-targeted and student-targeted attacks	In student attacks, the target scalarization switches to the student preference, while smoothing remains $\sigma = 0.5$ with train-time $m = 1$ and test-time $m = 100$

it. For the remaining cases, the student performs worse than all available teachers in only $15/116$ cases, which mainly appear in two scenarios. The first scenario is under Rand attacks. A possible explanation is that, when the two teachers have similar performance under the student preference, random state perturbations may introduce unstable perturbed trajectories and occasionally lead to performance collapse. The second scenario occurs when all available teachers already perform poorly under the target preference. In such cases, the teacher pool provides limited useful behavior for CAST to reuse.

J Full Effect of Smoothing on Student Robustness

This appendix reports the full smoothing comparison on all continuous and discrete environments. Tabs. 22 and 23 compare direct student aggregation without smoothing and smoothed aggregation. Across both continuous and discrete action spaces, smoothing improves the attacked return in $39/40$ configurations. The largest gain is 1534.50 (183.90% relative gain) on Walker2d under Rand at $\epsilon = 0.3$ in the continuous setting, and 16.41 on Deep Sea Treasure under PAAD at $\epsilon = 0.7$ in the discrete setting. The gains are especially clear under MAD, SARSA, and PAAD attacks.

These results show that the smoothing effect of certified teachers can be transferred to the CAST student through aggregation. The only decrease occurs on Deep Sea Treasure under PAAD at $\epsilon = 3.0$, where the unsmoothed student already has a very low cumulative reward. In this severe setting,

Table 15: Utility comparison among teachers, MORL-FB [51], and the smoothed student on continuous-control tasks under attacks.

Environment	ϵ	Model	Rand	MAD	SARSA	PAAD
Hopper $\sigma = 0.045$	0.2	teacher-118	850.26 ± 0.33	845.36 ± 0.55	845.77 ± 0.39	843.67 ± 0.32
		teacher-181	2203.54 ± 13.00	$\mathbf{2080.40 \pm 76.07}$	2191.54 ± 22.55	1750.68 ± 778.86
		teacher-811	$\mathbf{2274.16 \pm 9.73}$	1926.21 ± 553.95	$\mathbf{2240.51 \pm 12.92}$	$\mathbf{1980.93 \pm 606.06}$
		MORL-FB	1701.39 ± 346.02	521.80 ± 426.11	805.46 ± 193.71	706.44 ± 237.97
		stu. (CAST)	1786.17 ± 173.85	1584.06 ± 431.51	1125.19 ± 712.13	1838.71 ± 107.41
	0.75	teacher-118	839.00 ± 0.91	$\mathbf{771.20 \pm 11.37}$	748.47 ± 17.37	683.67 ± 353.06
		teacher-181	1505.44 ± 629.71	119.72 ± 90.22	158.93 ± 63.40	69.27 ± 4.63
		teacher-811	$\mathbf{1628.26 \pm 382.00}$	312.23 ± 114.25	260.73 ± 29.68	155.06 ± 47.92
		MORL-FB	151.60 ± 98.00	3.86 ± 9.76	378.99 ± 363.55	449.84 ± 368.01
		stu. (CAST)	1523.06 ± 541.11	622.98 ± 453.87	$\mathbf{797.91 \pm 564.53}$	$\mathbf{763.34 \pm 389.89}$
Walker2d $\sigma = 0.1$	0.3	teacher-91	2124.72 ± 64.78	1261.10 ± 502.47	842.02 ± 296.14	$\mathbf{1681.05 \pm 331.92}$
		teacher-19	993.12 ± 0.61	992.55 ± 0.68	989.30 ± 0.56	990.70 ± 1.02
		MORL-FB	1791.79 ± 600.77	-11.11 ± 93.01	553.44 ± 868.94	1113.00 ± 830.85
		stu. (CAST)	$\mathbf{2368.93 \pm 18.45}$	$\mathbf{2003.55 \pm 397.72}$	$\mathbf{1632.13 \pm 834.98}$	1263.13 ± 778.95
	1.0	teacher-91	$\mathbf{1263.68 \pm 447.34}$	-54.86 ± 54.72	-52.17 ± 27.83	64.48 ± 246.92
		teacher-19	985.60 ± 1.28	$\mathbf{437.32 \pm 299.79}$	$\mathbf{499.68 \pm 303.86}$	115.47 ± 20.47
		MORL-FB	-2.28 ± 48.55	-270.80 ± 107.11	-102.69 ± 16.58	-36.72 ± 28.51
		stu. (CAST)	1096.41 ± 719.40	117.67 ± 33.49	93.63 ± 31.35	$\mathbf{160.41 \pm 101.00}$
Humanoid $\sigma = 0.1$	1.0	teacher-19	53.96 ± 20.50	-3.74 ± 6.49	2.93 ± 7.23	13.38 ± 2.65
		teacher-91	73.05 ± 40.63	-6.09 ± 4.43	-7.01 ± 7.09	2.96 ± 7.81
		MORL-FB	$\mathbf{236.71 \pm 14.12}$	-18.88 ± 18.02	$\mathbf{135.02 \pm 33.86}$	9.17 ± 19.33
		stu. (CAST)	51.79 ± 9.16	$\mathbf{15.07 \pm 12.06}$	24.30 ± 12.08	$\mathbf{32.09 \pm 7.74}$
	3.0	teacher-19	27.22 ± 8.71	-2.09 ± 4.72	-1.27 ± 3.31	11.85 ± 10.23
		teacher-91	20.96 ± 7.09	-8.24 ± 4.90	-4.65 ± 2.31	5.93 ± 6.59
		MORL-FB	$\mathbf{140.81 \pm 63.64}$	-18.37 ± 13.56	-5.42 ± 21.20	-5.67 ± 10.39
		stu. (CAST)	38.30 ± 11.65	$\mathbf{14.08 \pm 8.68}$	$\mathbf{13.63 \pm 10.57}$	$\mathbf{31.06 \pm 7.04}$

the robustness gain introduced by smoothing may not be sufficient to offset the utility loss caused by locally perturbed teacher responses. Overall, these results suggest that CAST can transfer the robustness benefit of teacher smoothing to the student, improving empirical robustness under attacks.

K Full Action Certificate Diagnostics

In practical deployment, the smoothing parameters of CAST are usually selected according to target-preference utility, as discussed in Sec. F. This is a realistic choice because the student must first preserve target-preference utility after training-free adaptation. However, utility-oriented parameter selection does not fully characterize the robustness behavior of the resulting action certificates. We therefore conduct an additional diagnostic study to examine how the certification parameters affect the action-level bounds of CAST.

Continuous Action Certificate Diagnostics. Fig. 8 reports diagnostics for the continuous action certificate. According to Thm. 2, the bound is affected by the state smoothing scale σ, the perturbation radius ϵ, and the action smoothing scale σ_a. Since σ_a is an internal aggregation parameter and is fixed when constructing the student mixture, we focus on the external certification factors σ and ϵ. The clipping rate measures how often the adjusted quantile levels reach the boundary of $[0, 1]$. Clipping does not invalidate the certificate, but with $p = 0.5$, it causes the certified interval to cover the full action range, yielding a trivial bound. The interval width plots further show that larger perturbation radii generally widen the certified action intervals, while larger smoothing scales can improve the robustness of smoothed actions.

Discrete Action Certified Radius. Fig. 9 reports the discrete certified radius under different state smoothing scales σ. In Thm. 1, the certified radius is determined by the aggregated action gap and the smoothing induced Lipschitz constant. Since σ is the main certification parameter in the discrete setting, we evaluate the radius across different values of σ. The radius generally increases with σ, showing that stronger smoothing can enlarge the certified action invariance region.

Table 16: Utility comparison among teachers, MORL-FB [51], and the smoothed student on discrete-control tasks under attacks (bold values indicate the best reward in each comparison group, and underlined values indicate the second-best reward).

Environment	ϵ	Model	Rand	FGSM	PGD	PAAD
Deep Sea Treasure $\sigma = 0.5$	0.70	teacher-37	$\mathbf{2.46 \pm 0.91}$	$\mathbf{2.85 \pm 1.02}$	2.23 ± 1.56	$\mathbf{2.93 \pm 0.33}$
		teacher-73	2.35 ± 0.00	2.35 ± 0.00	$\mathbf{2.35 \pm 0.00}$	2.35 ± 0.00
		MORL-FB	-5.35 ± 12.88	-5.48 ± 12.82	-2.35 ± 11.33	-10.76 ± 14.24
		stu. (CAST)	2.39 ± 0.11	2.35 ± 0.00	$\mathbf{2.35 \pm 0.00}$	2.35 ± 0.00
	3.00	teacher-37	1.60 ± 1.33	-0.15 ± 0.00	-0.15 ± 0.00	-25.00 ± 0.00
		teacher-73	1.11 ± 1.67	2.67 ± 0.20	2.60 ± 0.00	$\mathbf{2.39 \pm 0.11}$
		MORL-FB	-4.36 ± 11.51	-6.56 ± 12.16	-2.95 ± 11.07	-25.00 ± 0.00
		stu. (CAST)	$\mathbf{2.01 \pm 1.12}$	$\mathbf{2.93 \pm 0.33}$	$\mathbf{2.89 \pm 0.37}$	-25.00 ± 0.00
Fruit-Tree $\sigma = 0.5$	0.30	teacher-511111	2.59 ± 0.25	3.30 ± 0.35	3.30 ± 0.35	3.14 ± 0.47
		teacher-151111	3.61 ± 0.00	$\mathbf{3.66 \pm 0.00}$	$\mathbf{3.66 \pm 0.00}$	3.61 ± 0.00
		teacher-115111	3.20 ± 0.29	3.39 ± 0.47	3.39 ± 0.47	2.90 ± 0.10
		teacher-111511	3.07 ± 0.09	3.18 ± 0.00	3.18 ± 0.00	3.18 ± 0.00
		teacher-111151	3.18 ± 0.45	3.27 ± 0.25	3.35 ± 0.29	3.15 ± 0.33
		teacher-111115	3.21 ± 0.25	3.49 ± 0.07	3.42 ± 0.31	3.42 ± 0.31
		MORL-FB	$\mathbf{3.80 \pm 0.05}$	$\mathbf{3.66 \pm 0.00}$	$\mathbf{3.66 \pm 0.00}$	3.15 ± 0.00
		stu. (CAST)	3.50 ± 0.22	3.39 ± 0.22	3.40 ± 0.27	$\mathbf{3.62 \pm 0.03}$
	1.00	teacher-511111	2.89 ± 0.48	3.08 ± 0.09	3.14 ± 0.00	3.57 ± 0.00
		teacher-151111	3.58 ± 0.11	3.49 ± 0.26	3.59 ± 0.16	3.28 ± 0.23
		teacher-115111	3.38 ± 0.33	$\mathbf{3.59 \pm 0.38}$	3.57 ± 0.17	$\mathbf{3.62 \pm 0.14}$
		teacher-111511	3.22 ± 0.22	3.21 ± 0.37	2.93 ± 0.43	2.90 ± 0.41
		teacher-111151	3.11 ± 0.43	3.57 ± 0.13	3.22 ± 0.35	3.43 ± 0.47
		teacher-111115	3.41 ± 0.24	3.36 ± 0.43	$\mathbf{3.61 \pm 0.34}$	3.14 ± 0.42
		MORL-FB	$\mathbf{3.61 \pm 0.05}$	3.55 ± 0.03	3.47 ± 0.05	2.69 ± 0.00
		stu. (CAST)	3.53 ± 0.14	3.56 ± 0.11	3.35 ± 0.40	3.50 ± 0.32

Together with Fig. 4 and Fig. 7, these diagnostics show the utility–certifiability trade-off of CAST. Parameters selected only for utility may not yield the largest certified radii or the narrowest certified action intervals, while parameters that improve certification can reduce target-preference utility when the smoothing becomes strong.

Table 17: Target preference sensitivity on Walker2d (bold values indicate the best reward in each comparison group, and underlined values indicate the second best reward).

Student Pref	Model	Rand ($\epsilon = 0.3$)	MAD ($\epsilon = 0.3$)	Rand ($\epsilon = 1.0$)	MAD ($\epsilon = 1.0$)
82	teacher-91	$\mathbf{3971.00 \pm 570.31}$	2244.89 ± 701.89	$\mathbf{2971.97 \pm 855.74}$	-8.86 ± 39.84
	teacher-19	992.47 ± 0.82	982.55 ± 1.62	982.17 ± 1.41	$\mathbf{550.02 \pm 277.27}$
	stu.	3746.28 ± 23.31	$\mathbf{2347.69 \pm 1160.92}$	1403.82 ± 1033.09	262.26 ± 240.32
73	teacher-91	2857.61 ± 862.16	1764.54 ± 563.99	$\mathbf{2296.86 \pm 885.03}$	-40.51 ± 34.12
	teacher-19	991.54 ± 0.66	977.99 ± 1.48	977.09 ± 1.34	$\mathbf{532.10 \pm 271.22}$
	stu.	$\mathbf{3211.75 \pm 128.24}$	$\mathbf{1939.50 \pm 1150.10}$	1237.91 ± 1096.95	215.47 ± 101.84
64	teacher-91	2292.46 ± 702.26	1284.20 ± 426.89	$\mathbf{1701.30 \pm 673.16}$	-72.17 ± 32.99
	teacher-19	990.70 ± 0.58	973.42 ± 1.38	971.81 ± 1.26	$\mathbf{514.18 \pm 265.30}$
	stu.	$\mathbf{2744.59 \pm 110.27}$	$\mathbf{1629.40 \pm 964.67}$	1042.09 ± 921.36	163.77 ± 81.36
46	teacher-91	1162.17 ± 383.47	323.50 ± 163.33	510.18 ± 253.39	-135.48 ± 44.49
	teacher-19	989.03 ± 0.44	964.28 ± 1.32	$\mathbf{961.27 \pm 1.33}$	$\mathbf{478.33 \pm 253.88}$
	stu.	$\mathbf{1810.26 \pm 74.80}$	$\mathbf{1009.21 \pm 593.82}$	650.46 ± 570.19	60.36 ± 41.53
37	teacher-91	597.03 ± 226.00	-156.84 ± 83.52	-85.38 ± 73.61	-167.13 ± 54.28
	teacher-19	988.20 ± 0.38	959.72 ± 1.37	$\mathbf{955.99 \pm 1.46}$	$\mathbf{460.41 \pm 248.40}$
	stu.	$\mathbf{1844.49 \pm 18.47}$	$\mathbf{1047.42 \pm 516.20}$	636.16 ± 463.83	68.38 ± 78.03
28	teacher-91	31.88 ± 80.58	-637.19 ± 161.22	-680.94 ± 192.57	-198.79 ± 65.27
	teacher-19	$\mathbf{987.36 \pm 0.33}$	$\mathbf{955.15 \pm 1.46}$	$\mathbf{950.72 \pm 1.65}$	$\mathbf{442.49 \pm 243.08}$
	stu.	875.93 ± 41.37	389.03 ± 223.19	258.84 ± 219.12	-43.04 ± 17.58

Table 18: Target preference sensitivity on Hopper (bold values indicate the best reward in each comparison group, and underlined values indicate the second best reward).

Student Pref	Model	Rand ($\epsilon = 0.2$)	MAD ($\epsilon = 0.2$)	Rand ($\epsilon = 0.75$)	MAD ($\epsilon = 0.75$)
722	teacher-811	$\mathbf{2876.81 \pm 6.66}$	$\mathbf{2099.57 \pm 928.31}$	1725.26 ± 724.46	468.64 ± 236.21
	teacher-181	1969.02 ± 468.78	1488.84 ± 760.30	$\mathbf{1903.98 \pm 454.81}$	160.94 ± 61.61
	teacher-118	935.39 ± 1.97	928.80 ± 1.19	927.70 ± 1.17	895.59 ± 4.50
	stu.	2084.66 ± 342.02	2028.89 ± 515.07	1814.42 ± 586.87	$\mathbf{1054.03 \pm 568.28}$
622	teacher-811	$\mathbf{2707.38 \pm 7.29}$	$\mathbf{1952.29 \pm 866.46}$	1587.58 ± 675.55	410.40 ± 207.21
	teacher-181	1956.65 ± 464.79	1463.24 ± 746.64	$\mathbf{1859.82 \pm 445.05}$	147.94 ± 57.95
	teacher-118	908.16 ± 2.93	898.64 ± 1.69	898.08 ± 1.17	855.29 ± 5.49
	stu.	1990.23 ± 315.26	1921.30 ± 487.68	1710.70 ± 552.02	$\mathbf{963.90 \pm 525.75}$
533	teacher-811	$\mathbf{2537.95 \pm 8.38}$	1805.01 ± 804.79	1449.91 ± 626.97	352.17 ± 178.34
	teacher-181	1944.28 ± 460.82	1437.64 ± 733.02	$\mathbf{1815.66 \pm 435.35}$	134.93 ± 54.37
	teacher-118	880.92 ± 3.98	868.45 ± 2.30	868.46 ± 1.24	815.00 ± 6.67
	stu.	1895.80 ± 288.87	$\mathbf{1813.70 \pm 460.44}$	1606.98 ± 517.25	$\mathbf{873.78 \pm 483.33}$
433	teacher-811	$\mathbf{2368.52 \pm 9.78}$	1657.74 ± 743.33	1312.24 ± 578.83	293.93 ± 149.64
	teacher-181	1931.91 ± 456.86	1412.03 ± 719.43	$\mathbf{1771.51 \pm 425.69}$	121.93 ± 50.90
	teacher-118	853.69 ± 5.06	838.28 ± 2.96	838.85 ± 1.38	774.70 ± 7.95
	stu.	1801.38 ± 262.96	$\mathbf{1706.11 \pm 433.40}$	1503.26 ± 482.59	$\mathbf{783.66 \pm 441.06}$

Table 19: Target preference sensitivity on Humanoid (bold values indicate the best reward in each comparison group, and underlined values indicate the second best reward).

Student Pref	Model	Rand ($\epsilon = 1.0$)	MAD ($\epsilon = 1.0$)	Rand ($\epsilon = 3.0$)	MAD ($\epsilon = 3.0$)
82	teacher-91	$\mathbf{184.35 \pm 99.03}$	25.01 ± 20.21	$\underline{83.00 \pm 25.18}$	18.97 ± 6.13
	teacher-19	$\underline{137.56 \pm 58.10}$	$\underline{29.86 \pm 11.48}$	$\mathbf{92.49 \pm 18.24}$	$\underline{26.70 \pm 8.42}$
	stu.	104.73 ± 14.67	$\mathbf{51.09 \pm 22.58}$	80.95 ± 18.51	$\mathbf{47.38 \pm 15.18}$
73	teacher-91	$\mathbf{147.25 \pm 79.54}$	14.64 ± 14.54	62.32 ± 18.88	9.90 ± 3.97
	teacher-19	$\underline{109.69 \pm 44.90}$	$\underline{18.66 \pm 9.10}$	$\mathbf{70.73 \pm 13.31}$	$\underline{17.10 \pm 6.83}$
	stu.	84.94 ± 11.58	$\mathbf{35.66 \pm 17.51}$	$\underline{64.45 \pm 15.21}$	$\mathbf{34.78 \pm 12.98}$
64	teacher-91	$\mathbf{110.15 \pm 60.07}$	4.28 ± 9.04	41.64 ± 12.73	0.83 ± 3.36
	teacher-19	$\underline{81.83 \pm 32.10}$	$\underline{7.46 \pm 7.28}$	$\mathbf{48.98 \pm 9.57}$	$\underline{7.51 \pm 5.51}$
	stu.	65.16 ± 9.28	$\mathbf{20.22 \pm 12.85}$	$\underline{47.94 \pm 12.13}$	$\mathbf{22.19 \pm 10.97}$
46	teacher-91	$\mathbf{35.95 \pm 21.31}$	-16.46 ± 5.03	0.28 ± 4.63	-17.31 ± 7.34
	teacher-19	$\underline{26.09 \pm 13.58}$	$\underline{-14.95 \pm 7.09}$	$\underline{5.47 \pm 11.41}$	$\underline{-11.69 \pm 4.71}$
	stu.	25.59 ± 9.45	$\mathbf{-10.65 \pm 8.19}$	$\mathbf{14.94 \pm 7.80}$	$\mathbf{-3.01 \pm 8.00}$
37	teacher-91	$\underline{-1.15 \pm 4.35}$	-26.83 ± 9.92	-20.40 ± 8.68	-26.38 ± 10.05
	teacher-19	-1.77 ± 17.99	$\underline{-26.15 \pm 8.80}$	$\underline{-16.29 \pm 15.95}$	$\underline{-21.28 \pm 5.49}$
	stu.	$\mathbf{5.81 \pm 11.86}$	$\mathbf{-26.08 \pm 10.52}$	$\mathbf{-1.57 \pm 7.66}$	$\mathbf{-15.60 \pm 7.49}$
28	teacher-91	-38.25 ± 18.57	$\mathbf{-37.19 \pm 15.47}$	-41.08 ± 14.57	-35.45 ± 12.86
	teacher-19	$\underline{-29.64 \pm 28.94}$	$\underline{-37.35 \pm 11.12}$	$\underline{-38.05 \pm 21.17}$	$\underline{-30.88 \pm 6.80}$
	stu.	$\mathbf{-13.98 \pm 15.00}$	-41.52 ± 14.69	$\mathbf{-18.07 \pm 9.15}$	$\mathbf{-28.20 \pm 7.85}$

Table 20: Target preference sensitivity on Deep Sea Treasure (bold values indicate the best reward in each comparison group, and underlined values indicate the second best reward).

Student Pref	Model	Rand ($\epsilon = 0.7$)	PGD ($\epsilon = 0.7$)	Rand ($\epsilon = 3.0$)	PGD ($\epsilon = 3.0$)
91	teacher-73	$\mathbf{19.33 \pm 0.29}$	$\underline{18.64 \pm 1.16}$	$\underline{16.50 \pm 5.38}$	$\underline{7.08 \pm 0.00}$
	teacher-37	6.60 ± 3.26	0.53 ± 0.00	6.01 ± 4.59	0.53 ± 0.00
	stu.	$\underline{19.24 \pm 0.39}$	$\mathbf{19.43 \pm 0.00}$	$\mathbf{18.18 \pm 1.96}$	$\mathbf{8.56 \pm 2.08}$
82	teacher-73	$\mathbf{15.16 \pm 0.00}$	$\underline{14.60 \pm 0.77}$	$\underline{13.30 \pm 1.50}$	$\underline{5.96 \pm 0.00}$
	teacher-37	5.51 ± 2.75	0.36 ± 0.00	3.44 ± 4.00	0.36 ± 0.00
	stu.	$\underline{15.03 \pm 0.26}$	$\mathbf{15.16 \pm 0.00}$	$\mathbf{13.56 \pm 1.70}$	$\mathbf{7.14 \pm 1.65}$
64	teacher-73	$\mathbf{6.62 \pm 0.01}$	$\underline{6.53 \pm 0.24}$	$\underline{4.11 \pm 1.75}$	$\underline{3.72 \pm 0.00}$
	teacher-37	$\underline{3.96 \pm 0.47}$	0.02 ± 0.00	2.27 ± 1.84	0.02 ± 0.00
	stu.	$\mathbf{6.62 \pm 0.00}$	$\mathbf{6.62 \pm 0.00}$	$\mathbf{5.28 \pm 1.08}$	$\mathbf{4.31 \pm 0.79}$
46	teacher-73	$\underline{-1.69 \pm 0.51}$	$\underline{-1.55 \pm 0.90}$	-3.57 ± 2.93	$\mathbf{1.48 \pm 0.00}$
	teacher-37	$\mathbf{0.98 \pm 0.85}$	$\mathbf{-0.32 \pm 0.00}$	$\mathbf{0.57 \pm 0.89}$	-0.32 ± 0.00
	stu.	-1.92 ± 0.00	-1.92 ± 0.00	$\underline{-3.17 \pm 3.73}$	$\underline{1.47 \pm 0.15}$
28	teacher-73	$\underline{-9.92 \pm 0.66}$	$\underline{-9.62 \pm 1.69}$	$\underline{-10.97 \pm 5.83}$	$\underline{-0.76 \pm 0.00}$
	teacher-37	$\mathbf{-0.94 \pm 0.38}$	$\mathbf{-0.66 \pm 0.00}$	$\mathbf{-1.75 \pm 1.75}$	$\mathbf{-0.66 \pm 0.00}$
	stu.	-10.46 ± 0.00	-10.46 ± 0.00	-11.50 ± 5.84	-1.36 ± 0.97
19	teacher-73	-14.73 ± 0.00	$\underline{-13.66 \pm 2.09}$	$\underline{-13.07 \pm 5.99}$	$\underline{-1.88 \pm 0.00}$
	teacher-37	$\mathbf{-1.67 \pm 0.42}$	$\mathbf{-0.83 \pm 0.00}$	$\mathbf{-3.26 \pm 2.69}$	$\mathbf{-0.83 \pm 0.00}$
	stu.	$\underline{-14.40 \pm 0.67}$	-14.73 ± 0.00	-15.44 ± 3.61	-2.78 ± 1.40

Table 21: Target preference sensitivity on Fruit Tree (bold values indicate the best reward in each comparison group, and underlined values indicate the second best reward).

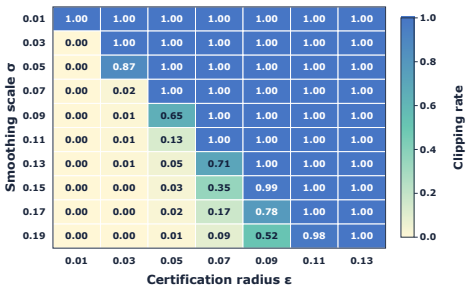
Student Pref	Model	Rand ($\epsilon = 0.3$)	PGD ($\epsilon = 0.3$)	Rand ($\epsilon = 1.0$)	PGD ($\epsilon = 1.0$)
$[0.9, 0.02 \times 5]$	teacher-511111	$\mathbf{6.97 \pm 2.16}$	2.99 ± 1.42	$\mathbf{5.20 \pm 2.12}$	1.61 ± 0.62
	teacher-151111	3.87 ± 0.22	$\underline{4.27 \pm 0.21}$	$\underline{4.21 \pm 1.61}$	$\underline{3.58 \pm 2.22}$
	teacher-115111	1.44 ± 0.55	3.06 ± 1.72	3.17 ± 2.30	3.21 ± 1.06
	teacher-111511	2.44 ± 0.19	4.20 ± 2.97	3.86 ± 2.02	$\mathbf{4.55 \pm 2.46}$
	teacher-111151	3.61 ± 2.46	2.56 ± 2.23	4.13 ± 2.87	3.26 ± 2.41
	teacher-111115	4.08 ± 1.74	$\mathbf{4.60 \pm 1.92}$	3.34 ± 1.51	2.39 ± 0.87
	stu.	$\underline{4.65 \pm 1.02}$	3.30 ± 1.40	4.15 ± 1.01	3.55 ± 1.64
$[0.8, 0.04 \times 5]$	teacher-511111	$\mathbf{6.46 \pm 1.75}$	3.02 ± 1.25	$\underline{4.40 \pm 1.29}$	1.83 ± 0.56
	teacher-151111	3.79 ± 0.17	$\underline{4.19 \pm 0.18}$	3.28 ± 1.48	$\underline{3.57 \pm 1.93}$
	teacher-115111	1.64 ± 0.96	3.11 ± 1.54	2.35 ± 1.37	3.24 ± 0.93
	teacher-111511	2.53 ± 0.15	4.03 ± 2.53	3.60 ± 1.61	$\mathbf{4.37 \pm 2.09}$
	teacher-111151	$\underline{5.37 \pm 2.26}$	2.66 ± 1.93	3.28 ± 1.84	3.24 ± 2.13
	teacher-111115	2.06 ± 0.82	$\mathbf{4.44 \pm 1.64}$	3.64 ± 1.23	2.55 ± 0.72
	stu.	3.79 ± 0.98	3.30 ± 1.24	$\mathbf{4.58 \pm 1.59}$	3.53 ± 1.41
$[0.7, 0.06 \times 5]$	teacher-511111	$\mathbf{5.85 \pm 1.46}$	3.06 ± 1.09	$\underline{4.45 \pm 1.49}$	2.05 ± 0.49
	teacher-151111	3.73 ± 0.11	$\underline{4.11 \pm 0.15}$	4.39 ± 1.39	$\underline{3.57 \pm 1.65}$
	teacher-115111	2.01 ± 0.80	3.16 ± 1.37	2.85 ± 1.93	3.27 ± 0.80
	teacher-111511	2.66 ± 0.10	3.87 ± 2.08	3.06 ± 0.82	$\mathbf{4.18 \pm 1.73}$
	teacher-111151	$\underline{4.50 \pm 2.30}$	2.75 ± 1.64	3.15 ± 1.43	3.22 ± 1.85
	teacher-111115	3.95 ± 0.82	$\mathbf{4.28 \pm 1.35}$	3.47 ± 1.21	2.71 ± 0.59
	stu.	4.28 ± 0.98	3.31 ± 1.07	$\mathbf{4.73 \pm 1.33}$	3.52 ± 1.19
$[0.6, 0.08 \times 5]$	teacher-511111	$\mathbf{5.40 \pm 1.00}$	3.10 ± 0.93	3.89 ± 1.21	2.27 ± 0.42
	teacher-151111	3.68 ± 0.00	$\underline{4.02 \pm 0.12}$	4.02 ± 1.17	$\underline{3.57 \pm 1.37}$
	teacher-115111	2.06 ± 0.33	3.21 ± 1.20	1.85 ± 0.21	3.30 ± 0.67
	teacher-111511	2.74 ± 0.07	3.70 ± 1.63	3.06 ± 1.04	$\mathbf{4.00 \pm 1.37}$
	teacher-111151	3.62 ± 1.29	2.84 ± 1.35	2.60 ± 1.31	3.21 ± 1.57
	teacher-111115	3.32 ± 0.98	$\mathbf{4.12 \pm 1.07}$	$\underline{4.15 \pm 0.47}$	2.87 ± 0.46
	stu.	$\underline{4.20 \pm 0.43}$	3.32 ± 0.90	$\mathbf{4.26 \pm 1.03}$	3.50 ± 0.97
$[0.4, 0.12 \times 5]$	teacher-511111	$\mathbf{4.24 \pm 0.39}$	3.18 ± 0.60	$\mathbf{4.19 \pm 0.38}$	2.72 ± 0.29
	teacher-151111	3.66 ± 0.06	$\mathbf{3.85 \pm 0.06}$	$\underline{3.85 \pm 0.30}$	$\underline{3.56 \pm 0.81}$
	teacher-115111	2.49 ± 0.14	3.32 ± 0.86	2.97 ± 0.73	3.37 ± 0.42
	teacher-111511	2.88 ± 0.00	3.37 ± 0.73	2.84 ± 0.14	$\mathbf{3.63 \pm 0.67}$
	teacher-111151	3.64 ± 0.66	3.03 ± 0.77	3.44 ± 0.46	3.17 ± 1.01
	teacher-111115	3.36 ± 0.51	$\underline{3.79 \pm 0.52}$	3.38 ± 0.64	3.18 ± 0.26
	stu.	$\underline{3.73 \pm 0.44}$	3.34 ± 0.58	3.69 ± 0.56	3.47 ± 0.56
$[0.3, 0.14 \times 5]$	teacher-511111	3.64 ± 0.08	3.22 ± 0.44	3.35 ± 0.35	2.94 ± 0.23
	teacher-151111	$\underline{3.67 \pm 0.06}$	$\mathbf{3.77 \pm 0.04}$	$\mathbf{3.68 \pm 0.31}$	$\mathbf{3.56 \pm 0.53}$
	teacher-115111	2.76 ± 0.17	3.37 ± 0.70	3.05 ± 0.53	3.40 ± 0.31
	teacher-111511	2.96 ± 0.04	3.20 ± 0.29	3.20 ± 0.34	3.45 ± 0.38
	teacher-111151	3.39 ± 0.38	3.13 ± 0.50	3.11 ± 0.42	3.15 ± 0.73
	teacher-111115	3.05 ± 0.47	$\underline{3.63 \pm 0.30}$	$\underline{3.48 \pm 0.37}$	3.34 ± 0.25
	stu.	$\mathbf{3.74 \pm 0.25}$	3.35 ± 0.43	$\mathbf{3.68 \pm 0.24}$	$\underline{3.46 \pm 0.40}$
$[0.2, 0.16 \times 5]$	teacher-511111	2.99 ± 0.31	3.26 ± 0.30	3.35 ± 0.33	3.16 ± 0.16
	teacher-151111	$\mathbf{3.64 \pm 0.03}$	$\mathbf{3.69 \pm 0.01}$	$\mathbf{3.59 \pm 0.13}$	$\mathbf{3.56 \pm 0.26}$
	teacher-115111	3.00 ± 0.17	3.42 ± 0.55	3.21 ± 0.33	3.43 ± 0.24
	teacher-111511	3.06 ± 0.08	3.04 ± 0.16	3.28 ± 0.32	3.26 ± 0.33
	teacher-111151	3.26 ± 0.36	3.22 ± 0.28	3.32 ± 0.34	3.13 ± 0.45
	teacher-111115	3.15 ± 0.34	$\underline{3.47 \pm 0.28}$	3.41 ± 0.29	$\underline{3.50 \pm 0.32}$
	stu.	$\underline{3.54 \pm 0.20}$	3.36 ± 0.31	$\underline{3.53 \pm 0.20}$	3.45 ± 0.32

Table 22: Robustness comparison between unsmoothed and smoothed student aggregation on continuous-control tasks under attacks, with all cumulative rewards evaluated using the student-preference scalarization and higher reward indicating stronger robustness under attack.

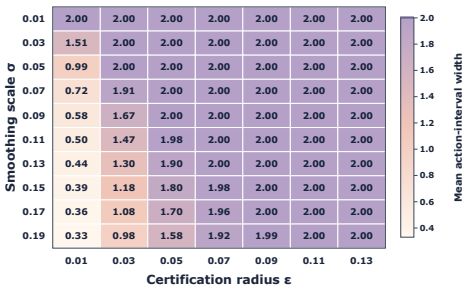
Environment	ϵ	Model	Rand	MAD
Hopper $\sigma = 0.045$	0.2	stu. (w/o smoothing)	818.13 ± 373.35	1186.13 ± 449.18
		stu. (w/ smoothing)	1786.17 ± 173.85	1584.06 ± 431.51
	0.75	stu. (w/o smoothing)	1101.57 ± 456.84	589.74 ± 213.57
		stu. (w/ smoothing)	1523.06 ± 541.11	622.98 ± 453.87
Walker2d $\sigma = 0.1$	0.3	stu. (w/o smoothing)	834.43 ± 341.97	682.41 ± 441.58
		stu. (w/ smoothing)	2368.93 ± 18.45	2003.55 ± 397.72
	1.0	stu. (w/o smoothing)	448.17 ± 288.63	-16.21 ± 124.91
		stu. (w/ smoothing)	1096.41 ± 719.40	117.67 ± 33.49
Humanoid $\sigma = 0.1$	1.0	stu. (w/o smoothing)	47.14 ± 9.38	-12.12 ± 7.26
		stu. (w/ smoothing)	51.79 ± 9.16	15.07 ± 12.06
	3.0	stu. (w/o smoothing)	22.69 ± 8.18	-5.70 ± 1.67
		stu. (w/ smoothing)	38.30 ± 11.65	14.08 ± 8.68
Environment	ϵ	Model	SARSA	PAAD
Hopper $\sigma = 0.045$	0.2	stu. (w/o smoothing)	996.97 ± 374.94	1135.12 ± 439.81
		stu. (w/ smoothing)	1125.19 ± 712.13	1838.71 ± 107.41
	0.75	stu. (w/o smoothing)	664.96 ± 400.26	657.70 ± 461.52
		stu. (w/ smoothing)	797.91 ± 564.53	763.34 ± 389.89
Walker2d $\sigma = 0.1$	0.3	stu. (w/o smoothing)	527.46 ± 547.19	840.21 ± 395.91
		stu. (w/ smoothing)	1632.13 ± 834.98	1263.13 ± 778.95
	1.0	stu. (w/o smoothing)	-10.38 ± 119.19	-63.68 ± 43.92
		stu. (w/ smoothing)	93.63 ± 31.35	160.41 ± 101.00
Humanoid $\sigma = 0.1$	1.0	stu. (w/o smoothing)	-2.41 ± 3.39	5.94 ± 7.80
		stu. (w/ smoothing)	24.30 ± 12.08	32.09 ± 7.74
	3.0	stu. (w/o smoothing)	0.20 ± 7.88	5.71 ± 6.56
		stu. (w/ smoothing)	13.63 ± 10.57	31.06 ± 7.04

Table 23: Robustness comparison between unsmoothed and smoothed student aggregation on discrete-control tasks under attacks, with all cumulative rewards evaluated using the student-preference scalarization and higher reward indicating stronger robustness under attack.

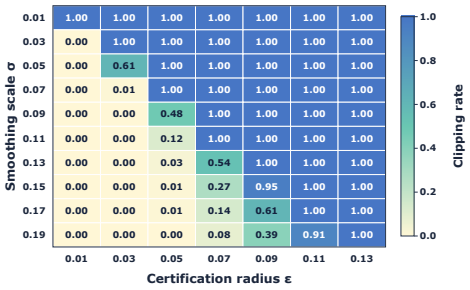
Environment	ϵ	Model	Rand	FGSM
Deep Sea Treasure $\sigma = 0.5$	0.70	stu. (w/o smoothing)	1.37 ± 2.26	-5.67 ± 12.67
		stu. (w/ smoothing)	2.39 ± 0.11	2.35 ± 0.00
	3.00	stu. (w/o smoothing)	1.96 ± 1.60	-0.58 ± 8.23
		stu. (w/ smoothing)	2.01 ± 1.12	2.93 ± 0.33
Fruit-Tree $\sigma = 0.5$	0.30	stu. (w/o smoothing)	3.34 ± 0.27	3.30 ± 0.37
		stu. (w/ smoothing)	3.50 ± 0.22	3.39 ± 0.22
	1.00	stu. (w/o smoothing)	3.40 ± 0.25	3.43 ± 0.25
		stu. (w/ smoothing)	3.53 ± 0.14	3.56 ± 0.11
Environment	ϵ	Model	PGD	PAAD
Deep Sea Treasure $\sigma = 0.5$	0.70	stu. (w/o smoothing)	-8.31 ± 13.64	-14.06 ± 13.40
		stu. (w/ smoothing)	2.35 ± 0.00	2.35 ± 0.00
	3.00	stu. (w/o smoothing)	-2.79 ± 11.11	-22.52 ± 7.46
		stu. (w/ smoothing)	2.89 ± 0.37	-25.00 ± 0.00
Fruit-Tree $\sigma = 0.5$	0.30	stu. (w/o smoothing)	3.28 ± 0.37	3.18 ± 0.30
		stu. (w/ smoothing)	3.40 ± 0.27	3.62 ± 0.03
	1.00	stu. (w/o smoothing)	3.33 ± 0.20	3.45 ± 0.35
		stu. (w/ smoothing)	3.35 ± 0.40	3.50 ± 0.32



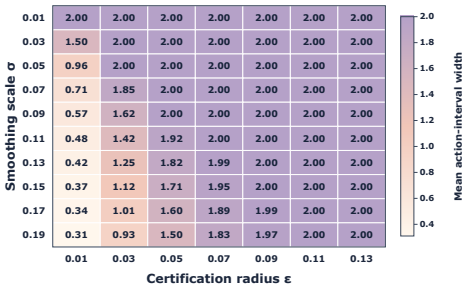
(a) Hopper clipping rate



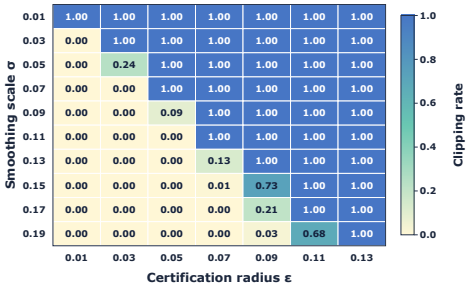
(b) Hopper certified interval



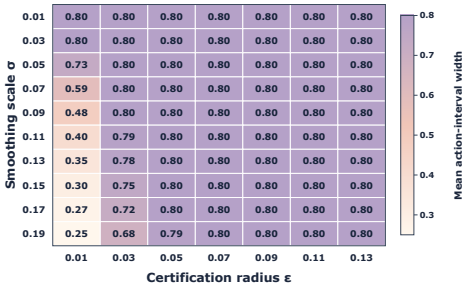
(c) Walker2d clipping rate



(d) Walker2d certified interval

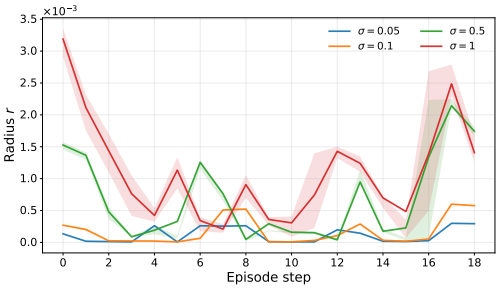


(e) Humanoid clipping rate

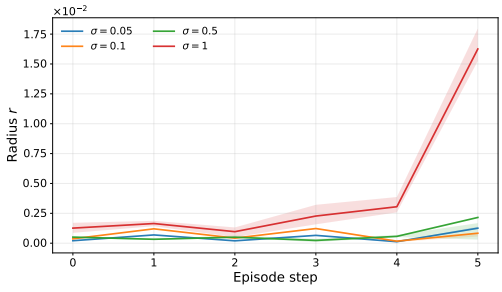


(f) Humanoid certified interval

Figure 8: Continuous action certificate diagnostics under different perturbation and smoothing settings (the first column reports the clipping rate of the adjusted quantile levels in Thm. 2, where a nonzero clipping rate indicates that some evaluated certified intervals cover the full action range, and the second column reports the certified interval).



(a) Deep Sea Treasure



(b) Fruit-Tree

Figure 9: Discrete action certified radius under different state-smoothing scales σ (larger smoothing scales generally increase the certified radius but can also reduce clean utility).

L Full Results of the Empirical Certification Evaluation

Tabs. 24 and 25 extend the representative action-certificate evaluation in the main text to all discrete- and continuous-action environments, respectively.

Discrete action certificate. Tab. 24 reports mean certified radii of 1.32×10^{-3} on Deep Sea Treasure and 1.26×10^{-3} on Fruit-Tree. All four attacks preserve the clean student action in every evaluation at $\epsilon \leq 10^{-2}$. Deep Sea Treasure maintains a 100% preservation rate through $\epsilon = 10^{-1}$, while Fruit-Tree retains rates between 90% and 98% at this perturbation magnitude. Action changes become more frequent at $\epsilon = 1$ and $\epsilon = 10$, as expected when the perturbations exceed the certified radii.

Continuous action certificate. Tab. 25 reports 100% joint coverage for every environment and perturbation magnitude, showing that all observed attacked actions lie within their coordinate-wise certified intervals. The certified interval width increases with ϵ in all three environments. On Hopper and Humanoid, the intervals reach the full action range at $\epsilon = 10^{-1}$, while on Walker2d they approach the full range at $\epsilon = 10^{-1}$ and cover it completely when $\epsilon \geq 1$. Therefore, when the perturbation magnitude ϵ is small relative to the smoothing scale σ, the certificates provide non-trivial bounds on the possible student actions, but these bounds become trivial once the certified intervals expand to the full action range.

M Full Certified Scalarized Cumulative Reward Results

Tabs. 26 and 27 report the numerical lower bounds for all environments. Since our experiments use multi-objective environments, Cor. 1 instantiates the vector-valued cumulative reward certificate in Thm. 3 under the target preference $\omega^\star$, and gives a scalarized cumulative reward certificate for the CAST student. To instantiate the environment smoothness terms required by Assump. 1, we estimate the scalarized reward Lipschitz constant $L_R^{\omega^\star}$ and the transition TV-Lipschitz constant L_P^{TV} from local transition data around evaluated trajectories, using empirical 0.99 quantile ratios for reward variation and a Gaussian surrogate for transition variation; details are in Appx. G. In all reported settings, the empirical rewards values are above the corresponding LBs, which is consistent with the certificate. These bounds are conservative because Cor. 1 gives a worst-case guarantee based on estimated reward and transition smoothness constants. The transition term can be large, especially in high-dimensional continuous-control tasks, and the factor $1/(1-\gamma)$ further amplifies local state-action variation when converting action certificates into cumulative reward guarantees. The bound can also become loose when the certified action interval is wide or when the certified action set contains multiple actions. Therefore, the LB values reported in this paper should be viewed as conservative certificate values. Despite this conservativeness, this result provides, to our knowledge, the first attempt to connect action-level certified bounds with cumulative-reward certification for an aggregated DRL student policy.

N Ablation Study

We conduct two sets of ablation studies to evaluate CAST's aggregation design: (1) examining the contributions of teacher weighting and nonlinear aggregation; and (2) comparing CAST with simpler alternatives for selecting and aggregating teacher outputs.

N.1 Contributions of Teacher Weighting and Nonlinear Aggregation

We analyze the contributions of teacher weighting and nonlinear aggregation in CAST by comparing four aggregation variants. Uniform Mean assigns equal weights to all teachers and combines their smoothed Q-values in discrete environments or smoothed action centers in continuous environments using an arithmetic mean. CAST-weighted Mean retains the margin-based teacher weighting of CAST but replaces nonlinear aggregation with an arithmetic mean. Uniform LSE and Uniform Quantile use equal teacher weights while retaining nonlinear aggregation, with LSE used for discrete action tasks and quantile aggregation for continuous action tasks, respectively. CAST (full) enables both margin-based teacher weighting and nonlinear aggregation.

Tab. 30 evaluates the effects of teacher weighting and nonlinear aggregation in continuous action tasks. CAST-weighted Mean improves over Uniform Mean in four of the six environment-budget

Table 24: Empirical validation of the discrete-action certificate on Deep Sea Treasure and Fruit-Tree with $\sigma = 1.0$.

Environment	Attack	Certified radius	$\epsilon = 10^{-4}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-1}$	$\epsilon = 1$	$\epsilon = 10$
Deep Sea Treasure	Rand	1.32×10^{-3}	100%	100%	100%	100%	88%	62%
	FGSM		100%	100%	100%	100%	40%	10%
	PGD		100%	100%	100%	100%	41%	31%
	PAAD		100%	100%	100%	100%	62%	35%
Fruit-Tree	Rand	1.26×10^{-3}	100%	100%	100%	98%	94%	72%
	FGSM		100%	100%	100%	90%	70%	30%
	PGD		100%	100%	100%	90%	59%	9%
	PAAD		100%	100%	100%	94%	82%	62%

Table 25: Empirical verification of the continuous action certificate under different perturbation radii with $\sigma = 0.09$ for Hopper and $\sigma = 0.19$ for Walker2d and Humanoid.

Environment	ϵ	Minimum certified interval width	Maximum action L_2 shift				Joint coverage across all attacks
			Rand	MAD	SARSA	PAAD	
Hopper	10^{-4}	0.008536	0.000443	0.000668	0.000501	0.000671	100.00%
	10^{-3}	0.085370	0.008807	0.007658	0.008773	0.008112	100.00%
	10^{-2}	0.861306	0.059725	0.097408	0.071540	0.076833	100.00%
	10^{-1}	3.464102	0.346464	0.539676	0.379732	0.379221	100.00%
	1	3.464102	1.411528	1.467051	0.946276	1.342919	100.00%
	10	3.464102	1.882620	2.601582	2.154228	2.330880	100.00%
Walker2d	10^{-4}	0.006355	0.000163	0.003719	0.000229	0.000301	100.00%
	10^{-3}	0.063547	0.008538	0.008388	0.005093	0.008288	100.00%
	10^{-2}	0.688459	0.021666	0.037412	0.030383	0.027429	100.00%
	10^{-1}	4.891142	0.133555	0.343670	0.175636	0.302095	100.00%
	1	4.898980	1.453948	1.919273	1.725271	1.632178	100.00%
	10	4.898980	2.050822	3.067277	2.615040	2.990106	100.00%
Humanoid	10^{-4}	0.009066	0.000982	0.001195	0.000043	0.000099	100.00%
	10^{-3}	0.090660	0.001207	0.002619	0.001244	0.002051	100.00%
	10^{-2}	0.890566	0.005630	0.017182	0.004875	0.008424	100.00%
	10^{-1}	3.298485	0.022946	0.144582	0.041984	0.115482	100.00%
	1	3.298485	0.233225	0.962989	0.384474	0.924709	100.00%
	10	3.298485	1.091841	1.826341	1.095250	1.758957	100.00%

settings, with only two small decreases of 1.90% and 0.13%. Uniform Quantile provides positive gains in all settings, ranging from 5.52% to 13.11%, indicating a more consistent benefit from nonlinear aggregation. CAST (full) also achieves positive gains in all settings. Overall, both teacher weighting and nonlinear aggregation show positive effects in the reported evaluations, while their relative contributions vary across environments and perturbation budgets.

Tab. 31 evaluates the effects of teacher weighting and nonlinear aggregation in discrete action tasks. CAST-weighted Mean improves over Uniform Mean in all environment-budget settings, with gains ranging from 0.97% to 13.80%. Uniform LSE also yields positive gains in all settings, ranging from 0.34% to 27.23%. These results indicate that both teacher weighting and nonlinear aggregation have positive effects when introduced individually. The effect of combining the two components varies across environments and perturbation budgets. In particular, at the larger perturbation budget, CAST (full) achieves the highest gains among the compared variants on both Deep Sea Treasure and Fruit Tree, reaching 48.11% and 2.23%, respectively.

Table 26: Certified quantities for the smoothed student on continuous-control tasks (the Rewards column reports the worst empirical reward of the smoothed student from Tab. 15, and we report the lower bound (LB), the scalarized $R_{\max}$, the reward Lipschitz estimate $\hat{L}_R^{\omega^\star}$, and the transition TV-Lipschitz estimate $\hat{L}_P^{\mathrm{TV}}$).

Environment	ϵ	Rewards	LB	$R_{\max}$	$\hat{L}_R^{\omega^\star}$	$\hat{L}_P^{\mathrm{TV}}$
Hopper	0.20	1125.19	866.48	3.39	0.54	2.48
$\sigma = 0.045$	0.75	622.98	-62.65			
Walker2d	0.30	1263.13	-4100.13	6.81	0.12	2.54
$\sigma = 0.1$	1.00	93.63	-6416.25			
Humanoid	1.00	15.07	-6301.69	2.11	0.00	8.39
$\sigma = 0.1$	3.00	13.63	-13301.15			

Table 27: Certified quantities for the smoothed student on discrete-control tasks using the finite-horizon lower bound (the Rewards column reports the worst empirical reward of the smoothed student from Tab. 16, and we report the lower bound (LB), the scalarized $R_{\max}$, the reward Lipschitz estimate $\hat{L}_R^{\omega^\star}$, and the transition TV-Lipschitz estimate $\hat{L}_P^{\mathrm{TV}}$).

Environment	ϵ	Rewards	LB	$R_{\max}$	$\hat{L}_R^{\omega^\star}$	$\hat{L}_P^{\mathrm{TV}}$
Deep Sea Treasure	0.7	2.35	-1690.15	11.35	11.8500	1.0000
$\sigma = 0.5$	3.0	-25.00	-3980.06			
Fruit-Tree	0.3	3.39	-447.12	3.95	3.9544	1.0000
$\sigma = 0.5$	1.0	3.35	-689.81			

N.2 Comparison with Teacher Selection and Aggregation Baselines

We compare CAST with simpler alternatives for selection and aggregation baselines to evaluate the effectiveness of its design choices. For continuous action tasks, we compare four aggregation modes: random teacher selection (RT), sample by weight aggregation (SBW), uniform weight aggregation with $p = 0.5$ (UWP05), and CAST. For discrete action tasks, we compare random teacher selection (RT), sample by weight aggregation (SBW), uniform weight greedy aggregation (UWG), and CAST. RT and SBW test whether simply selecting or sampling teachers is sufficient. UWP05 and UWG test whether uniform aggregation alone can explain the performance improvement. CAST further incorporates certified robustness signals into the aggregation process.

As shown in Tab. 28, for continuous action tasks, UWP05 and CAST clearly outperform RT and SBW in most settings. This shows that directly aggregating teacher outputs is the main source of improvement, while randomly selecting or sampling teachers is not sufficient. CAST is especially strong on Walker2d, where it is consistently better than UWP05 under all attacks and both perturbation budgets. This suggests that certified robustness signals provide better guidance beyond uniform aggregation. On Hopper and Humanoid, UWP05 is sometimes higher than CAST, but CAST remains competitive and performs better under several larger budget. Overall, the results show that direct aggregation is essential, and robustness guided aggregation further helps when teacher reliability varies across states.

As shown in Tab. 29, the discrete action results show a similar trend. On Deep Sea Treasure, UWG and CAST are much stronger than RT and SBW under the smaller perturbation budget, confirming that aggregating teacher action values is more reliable than selecting or sampling individual teachers. At $\epsilon = 3.0$, CAST performs best under PGD, suggesting that certified robustness signals can stabilize the aggregated action value profile and reduce harmful action switches. The main failure case is PAAD at $\epsilon = 3.0$, where both UWG and CAST reach the minimum reward. This indicates that aggregation can still fail when teacher signals become misleading under the target preference. On Fruit-Tree, all methods are closer, but CAST remains competitive and is often among the best performing variants. These results confirm that aggregation provides the main improvement, while certified robustness signals offer additional benefit when the teacher guidance is useful but uneven.

Table 28: Continuous attack-sm ablation results for four aggregation modes (entries report student scalarized cumulative reward means under Rand, MAD, SARSA, and PAAD attacks at two environment-specific perturbation budgets; RT = Random Teacher, SBW = Sample by Weight, UWP05 = Uniform Weight with $p = 0.5$; bold values indicate the best reward in each column, and underlined values indicate the second-best reward).

Perturbation	Mode	Rand	MAD	SARSA	PAAD
			Hopper		
$\epsilon_1 = 0.2$	RT	966.84 ± 480.87	852.45 ± 443.74	748.00 ± 404.84	763.50 ± 371.98
	SBW	1351.30 ± 527.93	789.11 ± 330.62	986.00 ± 718.75	1068.00 ± 628.75
	UWP05	$\mathbf{1889.62 \pm 0.00}$	$\mathbf{1624.30 \pm 323.72}$	$\mathbf{1370.80 \pm 716.42}$	$\underline{1833.64 \pm 103.75}$
	CAST	$\underline{1786.17 \pm 173.85}$	$\underline{1584.06 \pm 431.51}$	$\underline{1125.19 \pm 712.13}$	$\mathbf{1838.71 \pm 107.41}$
$\epsilon_2 = 0.75$	RT	841.59 ± 325.91	183.39 ± 108.60	227.85 ± 91.29	183.58 ± 58.77
	SBW	386.47 ± 204.09	229.30 ± 132.52	199.28 ± 47.25	247.11 ± 175.00
	UWP05	$\underline{1269.37 \pm 529.90}$	$\underline{506.61 \pm 332.93}$	$\underline{406.53 \pm 139.66}$	$\underline{381.72 \pm 225.40}$
	CAST	$\mathbf{1523.06 \pm 541.11}$	$\mathbf{622.98 \pm 453.87}$	$\mathbf{797.91 \pm 564.53}$	$\mathbf{763.34 \pm 389.89}$
			Walker2d		
$\epsilon_1 = 0.3$	RT	364.87 ± 432.94	91.96 ± 53.10	92.68 ± 53.27	372.51 ± 349.36
	SBW	588.86 ± 487.68	82.78 ± 102.51	62.43 ± 68.02	233.90 ± 246.54
	UWP05	$\underline{2107.89 \pm 523.15}$	$\underline{999.33 \pm 733.94}$	$\underline{587.20 \pm 221.81}$	$\underline{1231.70 \pm 761.47}$
	CAST	$\mathbf{2368.93 \pm 18.45}$	$\mathbf{2003.55 \pm 397.72}$	$\mathbf{1632.13 \pm 834.98}$	$\mathbf{1263.13 \pm 778.95}$
$\epsilon_2 = 1.0$	RT	172.67 ± 108.08	-87.48 ± 32.73	-35.99 ± 32.07	18.48 ± 20.08
	SBW	248.36 ± 171.82	-65.28 ± 37.31	-22.85 ± 16.52	3.52 ± 35.01
	UWP05	$\underline{743.81 \pm 654.64}$	$\underline{100.17 \pm 30.42}$	$\underline{86.97 \pm 31.54}$	$\underline{142.76 \pm 48.41}$
	CAST	$\mathbf{1096.41 \pm 719.40}$	$\mathbf{117.67 \pm 33.49}$	$\mathbf{93.63 \pm 31.35}$	$\mathbf{160.41 \pm 101.00}$
			Humanoid		
$\epsilon_1 = 1.0$	RT	38.07 ± 4.76	-2.86 ± 3.49	1.74 ± 5.52	8.77 ± 5.07
	SBW	42.45 ± 9.65	-7.05 ± 6.60	2.95 ± 5.52	8.66 ± 6.93
	UWP05	$\mathbf{65.82 \pm 6.15}$	$\mathbf{25.39 \pm 8.08}$	$\mathbf{30.10 \pm 7.92}$	$\mathbf{45.61 \pm 5.94}$
	CAST	$\underline{51.79 \pm 9.16}$	$\underline{15.07 \pm 12.06}$	$\underline{24.30 \pm 12.08}$	$\underline{32.09 \pm 7.74}$
$\epsilon_2 = 3.0$	RT	19.56 ± 10.46	-2.85 ± 6.77	-5.26 ± 2.40	11.64 ± 4.01
	SBW	20.57 ± 6.04	-4.78 ± 8.82	-3.69 ± 5.43	8.55 ± 7.79
	UWP05	$\mathbf{58.40 \pm 10.36}$	$\mathbf{19.31 \pm 8.08}$	$\mathbf{21.04 \pm 6.44}$	$\mathbf{51.99 \pm 5.13}$
	CAST	$\underline{38.30 \pm 11.65}$	$\underline{14.08 \pm 8.68}$	$\underline{13.63 \pm 10.57}$	$\underline{31.06 \pm 7.04}$

O Broader Impacts

Certified robust reinforcement learning has the potential to improve the reliability of sequential decision-making systems deployed under noisy or adversarial states. By providing formal guarantees within a specified perturbation set, methods such as CAST can help practitioners understand when a policy's action or cumulative reward is expected to remain stable. Such certificates may be useful in safety-relevant applications where empirical attack evaluation alone is insufficient to characterize worst-case behavior.

At the same time, certificates should not be interpreted beyond their stated assumptions. The guarantees in this work depend on the perturbation model, smoothing procedure, certified action set, and smoothness conditions used in the analysis. If the deployed environment violates these assumptions, for example through out-of-distribution dynamics, unmodeled sensor failures, or attacks outside the certified perturbation set, the certificate may no longer hold. Over-reliance on certificates in such settings could lead to misplaced confidence. Therefore, certified robustness should complement, rather than replace, empirical stress testing, monitoring, and domain-specific safety validation.

Table 29: Discrete ablation rewards on Deep Sea Treasure and Fruit Tree (RT = Random Teacher, SBW = Sample by Weight, UWG = Uniform Weight Greedy; each entry reports the student scalarized cumulative reward under a specific attack and perturbation radius; bold values indicate the best reward in each column; underlined values indicate the second-best reward).

Perturbation	Mode	Rand	FGSM	PGD	PAAD
Deep Sea Treasure					
$\epsilon_1 = 0.7$	RT	-0.50 ± 3.97	1.74 ± 2.07	0.55 ± 3.07	-1.90 ± 5.23
	SBW	1.52 ± 3.27	2.19 ± 2.65	$\mathbf{2.35 \pm 2.05}$	-1.53 ± 3.65
	UWG	2.35 ± 0.00	$\mathbf{2.35 \pm 0.00}$	$\mathbf{2.35 \pm 0.00}$	$\mathbf{2.35 \pm 0.00}$
	CAST	$\mathbf{2.38 \pm 0.10}$	$\mathbf{2.35 \pm 0.00}$	$\mathbf{2.35 \pm 0.00}$	$\mathbf{2.35 \pm 0.00}$
$\epsilon_2 = 3.0$	RT	1.03 ± 2.41	2.33 ± 0.82	1.34 ± 1.51	$\mathbf{-12.57 \pm 12.43}$
	SBW	1.69 ± 1.77	2.60 ± 0.00	1.53 ± 1.47	-20.03 ± 9.94
	UWG	$\mathbf{2.02 \pm 1.31}$	$\mathbf{2.92 \pm 0.33}$	1.91 ± 1.70	-25.00 ± 0.00
	CAST	2.01 ± 1.12	$\mathbf{2.92 \pm 0.33}$	$\mathbf{2.89 \pm 0.36}$	-25.00 ± 0.00
Fruit Tree					
$\epsilon_1 = 0.3$	RT	3.43 ± 0.24	3.40 ± 0.16	3.38 ± 0.23	3.47 ± 0.22
	SBW	$\mathbf{3.60 \pm 0.19}$	3.40 ± 0.26	$\mathbf{3.44 \pm 0.28}$	3.59 ± 0.19
	UWG	3.57 ± 0.08	$\mathbf{3.49 \pm 0.22}$	3.40 ± 0.27	3.58 ± 0.06
	CAST	3.50 ± 0.21	3.39 ± 0.22	3.40 ± 0.27	$\mathbf{3.62 \pm 0.03}$
$\epsilon_2 = 1.0$	RT	3.40 ± 0.18	3.47 ± 0.26	$\mathbf{3.46 \pm 0.17}$	3.40 ± 0.33
	SBW	$\mathbf{3.53 \pm 0.38}$	3.37 ± 0.40	3.30 ± 0.43	$\mathbf{3.58 \pm 0.24}$
	UWG	3.46 ± 0.14	3.54 ± 0.21	3.27 ± 0.43	3.33 ± 0.40
	CAST	$\mathbf{3.53 \pm 0.14}$	$\mathbf{3.56 \pm 0.11}$	3.35 ± 0.40	3.50 ± 0.32

Table 30: Ablation of CAST's weighting and aggregation components (Uniform Mean and CAST-weighted Mean use arithmetic averaging with equal and CAST weights, respectively. Uniform Quantile uses equal weights with quantile aggregation. The symbol pairs in the column headers indicate whether margin weighting and quantile aggregation are enabled, respectively. Avg. gain is the arithmetic mean of the percentage improvements over Uniform Mean across the four attacks).

Perturbation	Attack	Uniform Mean (×, ×)	CAST-weighted Mean (✓, ×)	Uniform Quantile (×, ✓)	CAST (full) (✓, ✓)
			Hopper		
$\epsilon_1 = 0.2$	Rand	1621.63 ± 63.50	1603.20 ± 355.53	1811.97 ± 204.94	1786.17 ± 173.85
	MAD	1530.03 ± 85.30	1647.61 ± 168.23	1655.14 ± 247.05	1584.06 ± 431.51
	SARSA	1383.26 ± 7.66	1000.41 ± 678.67	1436.40 ± 654.31	1125.19 ± 712.13
	PAAD	1506.24 ± 335.06	1709.96 ± 175.23	1787.29 ± 183.57	1838.71 ± 107.41
	Avg. gain	0.00%	-1.90%	$+10.60\%$	$+4.27\%$
$\epsilon_2 = 0.75$	Rand	1535.64 ± 77.45	1594.77 ± 204.71	1716.20 ± 105.54	1523.06 ± 541.11
	MAD	331.08 ± 149.84	516.22 ± 363.96	501.80 ± 409.00	622.98 ± 453.87
	SARSA	429.63 ± 123.34	347.54 ± 227.92	474.70 ± 282.04	797.91 ± 564.53
	PAAD	619.64 ± 240.49	565.67 ± 425.95	396.10 ± 171.36	763.34 ± 389.89
	Avg. gain	0.00%	$+7.99\%$	$+9.43\%$	$+49.06\%$
			Walker2d		
$\epsilon_1 = 0.3$	Rand	2271.93 ± 39.50	2309.94 ± 17.39	2271.84 ± 37.54	2368.93 ± 18.45
	MAD	589.53 ± 344.97	1212.14 ± 670.49	731.66 ± 241.05	2003.55 ± 397.72
	SARSA	701.70 ± 373.22	678.45 ± 211.69	642.71 ± 320.07	1632.13 ± 834.98
	PAAD	1174.95 ± 779.66	1267.52 ± 801.77	1249.77 ± 813.49	1263.13 ± 778.95
	Avg. gain	0.00%	$+27.96\%$	$+5.52\%$	$+96.06\%$
$\epsilon_2 = 1.0$	Rand	921.56 ± 650.52	1016.34 ± 683.96	927.24 ± 665.17	1096.41 ± 719.40
	MAD	91.69 ± 12.24	90.26 ± 31.24	94.12 ± 40.81	117.67 ± 33.49
	SARSA	70.55 ± 28.47	84.10 ± 26.52	83.68 ± 24.46	93.63 ± 31.35
	PAAD	134.59 ± 45.62	134.59 ± 54.96	139.35 ± 52.88	160.41 ± 101.00
	Avg. gain	0.00%	$+6.98\%$	$+6.35\%$	$+24.80\%$
			Humanoid		
$\epsilon_1 = 1.0$	Rand	61.27 ± 0.91	60.64 ± 1.96	61.27 ± 0.91	51.79 ± 9.16
	MAD	11.09 ± 2.84	10.19 ± 6.17	17.21 ± 2.76	15.07 ± 12.06
	SARSA	19.62 ± 0.30	22.43 ± 3.31	19.79 ± 1.14	24.30 ± 12.08
	PAAD	39.76 ± 3.24	37.49 ± 2.43	38.32 ± 3.22	32.09 ± 7.74
	Avg. gain	0.00%	-0.13%	$+13.11\%$	$+6.25\%$
$\epsilon_2 = 3.0$	Rand	46.53 ± 0.49	48.23 ± 2.37	46.53 ± 0.49	38.30 ± 11.65
	MAD	3.19 ± 8.06	11.48 ± 3.38	5.52 ± 2.90	14.08 ± 8.68
	SARSA	16.07 ± 2.06	15.88 ± 3.00	13.55 ± 0.76	13.63 ± 10.57
	PAAD	48.66 ± 0.47	37.27 ± 2.41	44.39 ± 3.40	31.06 ± 7.04
	Avg. gain	0.00%	$+59.62\%$	$+12.07\%$	$+67.90\%$

Table 31: Ablation of CAST's weighting and aggregation components on Deep Sea Treasure and Fruit Tree (Uniform Mean and CAST-weighted Mean use arithmetic averaging with equal and CAST weights, respectively. Uniform LSE uses equal weights with log-sum-exp aggregation. The symbol pairs in the column headers indicate whether margin weighting and LSE aggregation are enabled, respectively. Avg. gain is the arithmetic mean of the percentage improvements over Uniform Mean across the four attacks).

Perturbation	Attack	Uniform Mean ($\times$, $\times$)	CAST-weighted Mean ($\checkmark$, $\times$)	Uniform LSE ($\times$, $\checkmark$)	CAST (full) ($\checkmark$, $\checkmark$)
Deep Sea Treasure					
$\epsilon_1 = 0.7$	Rand	2.35 ± 0.00	2.45 ± 0.16	2.38 ± 0.10	2.38 ± 0.10
	FGSM	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00
	PGD	2.35 ± 0.00	2.38 ± 0.10	2.38 ± 0.10	2.35 ± 0.00
	PAAD	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00	2.35 ± 0.00
	Avg. gain	0.00%	$+1.49\%$	$+0.74\%$	$+0.37\%$
$\epsilon_2 = 3.0$	Rand	1.41 ± 2.18	$1.95 \pm 1.62^{\ddagger}$	1.73 ± 1.42	2.01 ± 1.12
	FGSM	2.60 ± 0.00	2.60 ± 0.00	2.73 ± 0.26	2.93 ± 0.33
	PGD	1.22 ± 1.82	1.42 ± 1.59	2.21 ± 1.58	2.89 ± 0.36
	PAAD	-25.00 ± 0.00	-25.00 ± 0.00	-25.00 ± 0.00	-25.00 ± 0.00
	Avg. gain	0.00%	$+13.80\%$	$+27.23\%$	$+48.11\%$
Fruit Tree					
$\epsilon_1 = 0.3$	Rand	3.53 ± 0.23	3.50 ± 0.21	3.51 ± 0.17	3.50 ± 0.21
	FGSM	3.46 ± 0.21	3.45 ± 0.23	3.49 ± 0.22	3.39 ± 0.22
	PGD	3.33 ± 0.21	3.36 ± 0.32	3.46 ± 0.24	3.40 ± 0.27
	PAAD	3.37 ± 0.34	3.57 ± 0.13	3.58 ± 0.06	3.62 ± 0.03
	Avg. gain	0.00%	$+1.30\%$	$+2.54\%$	$+1.49\%$
$\epsilon_2 = 1.0$	Rand	3.49 ± 0.25	3.37 ± 0.21	3.43 ± 0.13	3.53 ± 0.14
	FGSM	3.33 ± 0.34	3.43 ± 0.34	3.53 ± 0.21	3.56 ± 0.11
	PGD	3.33 ± 0.35	3.50 ± 0.16	3.45 ± 0.26	3.35 ± 0.40
	PAAD	3.50 ± 0.21	3.48 ± 0.35	3.27 ± 0.48	3.50 ± 0.32
	Avg. gain	0.00%	$+0.97\%$	$+0.34\%$	$+2.23\%$

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by the method in Section 4, the certification results in Section 5, and the experiments in Section 6. Full experimental results and sensitivity analyses are provided in Appendices H - N.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: In Appendix O, we discuss the limitations of our certificates, including their dependence on the stated perturbation model and smoothness assumptions, and their potential invalidity when these assumptions are violated.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The theoretical assumptions and statements are given in Sections 3 and 5, including the environment smoothness assumption in Section 5.2. Complete proofs are provided in Appendices B - D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The main reproducibility information is provided in Section 6, including environments, preferences, certification backbones, and attack settings. Parameter selection, full result tables, runtime, computational resources, and training details are provided in Appendices F, H, I, and G.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide publicly available code and instructions for reproducing the main experimental results at `https://github.com/Ryndalf/CAST`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies the main experimental settings in Section 6, including environments, teacher and student preference construction, certification backbones, and attack settings. Hyperparameter and parameter-selection details are provided in Appendix F, and training details are provided in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main and appendix result tables report mean performance together with standard deviations. These statistics are shown in Section 6 and in the full result tables in Appendices H, J, I, and N.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper reports runtime and computational resources in Appendix G.1. Additional training details for continuous-action and discrete-action experiments are provided in Appendices G.2 and G.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The experiments use simulated reinforcement learning environments described in Section 6, and the work does not involve human subjects, private data, or sensitive personal information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses broader impacts in Appendix O. It covers potential benefits for reliable sequential decision-making and potential risks from over-interpreting certificates beyond the stated perturbation model and smoothness assumptions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets that require additional safeguards for responsible use. The experiments are conducted in the simulated reinforcement learning environments described in Section 6.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper credits the existing environments, certification methods, and attack methods in Section 6. The licenses and terms of use for code-related assets will be documented in the released code README.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not release new datasets, models, or benchmark assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. The empirical study uses simulated reinforcement learning environments described in Section 6.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects or crowdsourcing experiments, so IRB approval or equivalent review is not applicable. The experiments are conducted in simulated reinforcement learning environments described in Section 6.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing, editing, and formatting assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.