DAGA: Distributional Actor-Critic with Action-Gradient Alignment and Stability-Triggered Guidance

Anonymous authors
Paper under double-blind review

Abstract

Multi-critic actor–critic methods improve continuous-control learning by aggregating critic values or return distributions, yet deterministic actors are updated through critic action gradients. Critics can therefore agree in value while inducing different policy-update directions, a phenomenon we call value–gradient mismatch. We propose DAGA, a Gaussian multi-critic framework combining Stability-Triggered Guidance (STG), which adapts actor guidance using distribution-level reliability, with Action-Gradient Policy Alignment (AGPA), which selectively regularizes actor-relevant gradient directions when excessive dispersion coexists with a coherent consensus. We characterize value–gradient decoupling and both mechanisms locally. Experiments evaluate mismatch prevalence, component contributions, cross-task performance, and sensitivity to intervention strength and ensemble size.

1 Introduction

Multi-critic actor–critic methods improve continuous-control learning through mechanisms such as conservative targets, ensemble subsampling, distributional truncation, and Gaussian critic fusion (Fujimoto et al., 2018; Chen et al., 2021; Kuznetsov et al., 2020; Valencia et al., 2025). These mechanisms primarily operate on predicted values or return distributions. A deterministic actor, however, is updated through critic derivatives with respect to actions. Critics may therefore agree at the current policy action while inducing different or opposing update directions. We call this pointwise decoupling value–gradient mismatch. Although critic action gradients are important optimization objects (D'Oro & Jaśkowski, 2020; Gurumurthy et al., 2023; Lee et al., 2026), their directional agreement across critics estimating the same objective has received less attention than value-level critic aggregation.

Critic disagreement is not inherently harmful: ensemble diversity can be useful and is deliberately encouraged in some offline settings (An et al., 2021). The relevant question is whether excessive actor-relevant directional disagreement is associated with unstable learning and whether intervention can remain selective.

We address this question with DAGA (Distributional Actor-Critic with Action-Gradient Alignment), built on a Gaussian multi-critic backbone. DAGA separates distribution-level reliability from action-gradient geometry. Stability-Triggered Guidance (STG) adapts the critic weights used for actor guidance using critic-specific distributional scale and ensemble-relative disagreement. Action-Gradient Policy Alignment (AGPA) selectively regularizes actor-relevant directions when dispersion exceeds a dead zone and a coherent consensus remains, preserving a non-intervention region rather than enforcing unconditional alignment. Our contributions are:

- We formulate multi-critic value–gradient mismatch and introduce diagnostics separating distribution-level disagreement from actor-relevant gradient geometry.
- We propose DAGA, combining disagreement-aware STG with gated action-gradient alignment through AGPA while retaining a fixed distributional backbone.
- We characterize the mechanisms locally and evaluate mismatch prevalence, component contributions, cross-task performance, and sensitivity.

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

2 Related Work

Multi-critic methods typically manipulate critic values or return distributions. TD3 uses clipped double-Q targets, REDQ performs randomized in-target minimization over an ensemble, TQC truncates upper distributional quantiles, and CTD4 combines Gaussian critic distributions through Kalman-style fusion (Fujimoto et al., 2018; Chen et al., 2021; Kuznetsov et al., 2020; Valencia et al., 2025). Distributional RL more broadly models return distributions (Bellemare et al., 2017; Dabney et al., 2018; Barth-Maron et al., 2018). Adaptive ensemble aggregation has also used disagreement to modify value-level fusion (Werge et al., 2025). DAGA instead uses a fixed Gaussian aggregation backbone and introduces ensemble-relative reliability specifically for actor guidance, without treating return variance as calibrated epistemic uncertainty. Prior work has shown that critic derivatives contain information not captured by scalar value fitting alone. MAGE explicitly learns action-value gradients using model-based information (D'Oro & Jaśkowski, 2020), while CGAC studies critic-action-gradient policy updates and instability arising from biased critic gradients in an on-policy setting (Gurumurthy et al., 2023). PAVE regularizes the critic's Q-gradient field to improve policy smoothness (Lee et al., 2026), and recent diffusion/flow-policy methods also exploit Q-gradients for policy optimization (Fang et al., 2025; Li & Levine, 2026). The vMFER resamples experience using directional uncertainty for actor–critic learning (Zhu et al., 2024), while EDAC encourages normalized action-gradient diversity for offline pessimism (An et al., 2021). In contrast, DAGA studies directional agreement across simultaneously trained critics in online deterministic control and selectively increases concentration only when a coherent consensus remains. Its intervention is through critic-specific actor guidance (STG) and gated directional regularization (AGPA), rather than experience resampling. It therefore targets the decoupling between critic value/distribution agreement and actor-facing gradient geometry. PCGrad and CAGrad resolve conflicting parameter gradients from distinct task objectives (Yu et al., 2020; Liu et al., 2021). DAGA instead analyzes critic-output gradients with respect to actions for critics trained on the same control objective, multi-critic disagreement is therefore not formulated as a multi-task optimization problem.

3 Methodology

3.1 Problem Setup and Value-Gradient Mismatch

Figure 1 summarizes DAGA. We distinguish the precision-fused distribution used by the backbone from a moment-matched ensemble reference used for reliability assessment, then define action-gradient geometry and the STG and AGPA mechanisms. We consider a discounted Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ and $\gamma \in (0, 1)$. A deterministic actor $a = \pi_\theta(s)$ is trained together with N separately parameterized critics. Critic i, parameterized by ϕ_i, models the conditional return distribution as:

$$Z_{\phi_i}(s, a) \sim \mathcal{N}\left(\mu_i(s, a), v_i(s, a)\right) \tag{1}$$

where μ_i and v_i denote the predicted return mean and variance, respectively. We use v_i as a critic-specific distributional scale, without assuming that it is a calibrated estimate of epistemic uncertainty. Since $Q_i(s, a) := \mathbb{E}[Z_{\phi_i}(s, a)] = \mu_i(s, a)$, the actor-facing action gradient of critic i is:

$$g_i(s) = \nabla_a \mu_i(s, a)|_{a=\pi_\theta(s)} \tag{2}$$

Let $J_\theta(s) = \partial \pi_\theta(s)/\partial \theta$ denote the policy Jacobian. Holding the critic parameters and replay-state distribution fixed, the corresponding critic-specific contribution to the actor surrogate gradient is:

$$p_i(s) = \left[J_\theta(s)^\top g_i(s)\right] \tag{3}$$

DAGA analyzes $g_i(s)$ rather than $p_i(s)$ because it isolates the critic-specific component of the update and lies in the typically lower-dimensional action space. Since the policy Jacobian need not preserve directional disagreement, we treat action-gradient geometry as a critic-level diagnostic and empirically examine its propagation to actor-parameter updates. Two

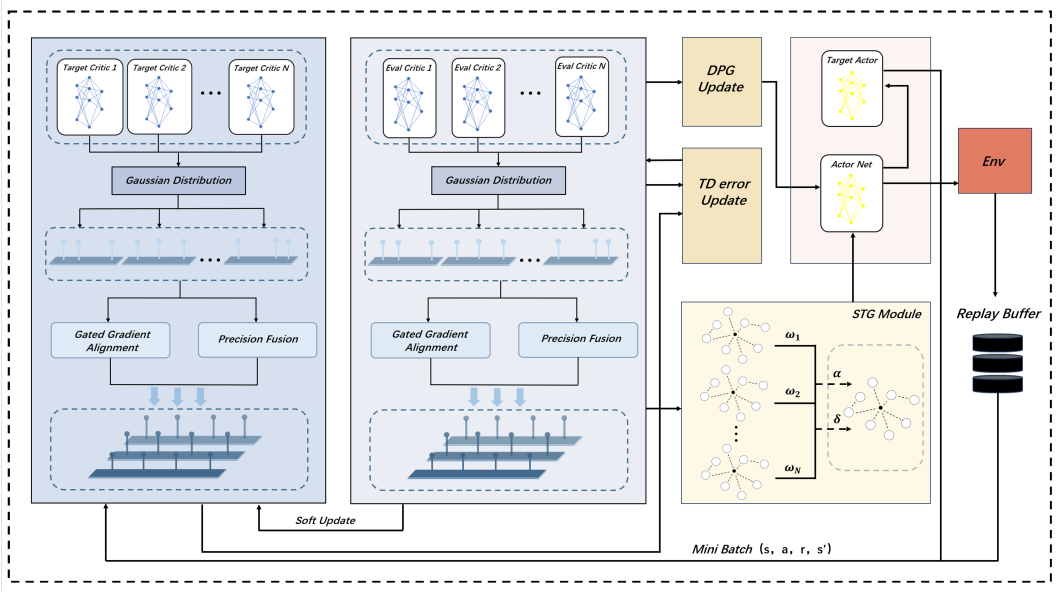

Figure 1: Overview of DAGA. Precision fusion provides value-level aggregation, STG adapts critic contributions to actor guidance, and AGPA selectively regularizes actor-facing action-gradient geometry.

critics may satisfy:

$$\mu_i(s, \pi_\theta(s)) \approx \mu_j(s, \pi_\theta(s)) \tag{4}$$

while $g_i(s)$ and $g_j(s)$ point in different directions. We refer to this pointwise decoupling between value agreement and action-gradient agreement as value–gradient mismatch. This definition is diagnostic and does not assume that critic disagreement is inherently harmful.

3.2 Distributional Aggregation and Reliability Reference

For each critic i, we define the stabilized precision and its normalized weight as:

$$\tau_i(s, a) = \frac{1}{v_i(s, a) + \epsilon_v} \tag{5}$$

$$\alpha_i(s, a) = \frac{\tau_i(s, a)}{\sum_{j=1}^{N} \tau_j(s, a)} \tag{6}$$

The precision-fused Gaussian is parameterized by:

$$\mu_F(s, a) = \sum_{i=1}^{N} \alpha_i(s, a)\mu_i(s, a) \tag{7}$$

$$v_F(s, a) = \left(\sum_{i=1}^{N} \tau_i(s, a) \right)^{-1} \tag{8}$$

We adopt this rule as the fixed value-level aggregation mechanism of the Gaussian multi-critic backbone rather than as a contribution of DAGA. It is used both for Bellman-target construction and as the baseline actor weighting. Its conditional statistical interpretation is analyzed in Appendix A. Using the precision-fused distribution itself as a disagreement reference would mechanically contract its variance with ensemble size. We therefore construct

a separate Gaussian reference by matching the first two moments of the equally weighted critic mixture:

$$\bar{Z}(s,a) = \mathcal{N}\left(\bar{\mu}(s,a), \bar{v}(s,a)\right) \tag{9}$$

$$\bar{\mu}(s,a) = \frac{1}{N}\sum_{i=1}^{N} \mu_i(s,a) \tag{10}$$

$$\bar{v}(s,a) = \frac{1}{N}\sum_{i=1}^{N} \left[v_i(s,a) + \big(\mu_i(s,a) - \bar{\mu}(s,a)\big)^2\right] \tag{11}$$

The two terms in $\bar{v}$ represent the average within-critic return variance and the between-critic dispersion of predicted means, respectively. If all critics are identical, $Z_{\phi_i} = \mathcal{N}(\mu, v)$, then $\bar{Z} = \mathcal{N}(\mu, v)$ independently of N, whereas precision fusion gives $v_F = (v + \epsilon_v)/N$. We quantify deviation of critics from the ensemble reference using the symmetric Gaussian KL divergence:

$$\delta_i(s,a) = \frac{1}{2}\left[D_{\mathrm{KL}}\left(Z_{\phi_i}(s,a) \parallel \bar{Z}(s,a)\right) + D_{\mathrm{KL}}\left(\bar{Z}(s,a) \parallel Z_{\phi_i}(s,a)\right)\right] \tag{12}$$

and define the mean disagreement as:

$$\bar{\delta}(s,a) = \frac{1}{N}\sum_{i=1}^{N} \delta_i(s,a) \tag{13}$$

We interpret δ_i as an ensemble-relative disagreement score rather than as a calibrated probability that critic i is erroneous.

3.3 Action-Gradient Geometry

For each replay state s, define the set of critics with numerically valid action-gradient directions and its cardinality as:

$$\mathcal{I}(s) = \{i : \|g_i(s)\|_2 > \epsilon_g\} \tag{14}$$

and $M(s) = |\mathcal{I}(s)|$. For $i \in \mathcal{I}(s)$, the normalized action-gradient direction is:

$$u_i(s) = \frac{g_i(s)}{\|g_i(s)\|_2} \tag{15}$$

Directional diagnostics are evaluated only when $M(s) \geq 2$. The pairwise cosine similarity and conflict ratio are:

$$C_{\mathrm{mean}}(s) = \frac{2}{M(s)\big(M(s)-1\big)} \sum_{\substack{i<j \\ i,j\in\mathcal{I}(s)}} u_i(s)^\top u_j(s) \tag{16}$$

$$R_{\mathrm{conf}}(s) = \frac{2}{M(s)\big(M(s)-1\big)} \sum_{\substack{i<j \\ i,j\in\mathcal{I}(s)}} \mathbf{1}\left\{u_i(s)^\top u_j(s) < 0\right\} \tag{17}$$

We report their replay-state averages. For any nonnegative weights $w_i(s)$ satisfying $\sum_{i\in\mathcal{I}(s)} w_i(s) = 1$, define:

$$m_w(s) = \sum_{i\in\mathcal{I}(s)} w_i(s) u_i(s) \tag{18}$$

$$R_w(s) = \|m_w(s)\|_2 \tag{19}$$

and

$$A_w(s) = \sum_{i \in \mathcal{I}(s)} w_i(s) \|u_i(s) - m_w(s)\|_2^2 \tag{20}$$

Since $\|u_i\|_2 = 1$, it follows that:

$$A_w(s) = 1 - R_w(s)^2 \tag{21}$$

Thus R_w and A_w equivalently characterize directional concentration. We additionally track the average gradient magnitude:

$$G_{\text{norm}} = \mathbb{E}_{s \sim \mathcal{D}} \left[\frac{1}{N} \sum_{i=1}^{N} \|g_i(s)\|_2 \right] \tag{22}$$

which distinguishes increased directional agreement from the degenerate case in which all critic gradients merely shrink toward zero.

3.4 Distributional Critic Learning

For a transition (s, a, r, s', d), where d denotes the terminal flag, the target action is:

$$a' = \text{clip}_{\mathcal{A}}(\pi_{\bar{\theta}}(s') + \epsilon_{\text{tps}}) \tag{23}$$

where $\bar{\theta}$ denotes the target-actor parameters and ϵ_{tps} denotes clipped zero-mean Gaussian target-policy-smoothing noise with standard deviation σ_{tps}. Applying the same precision-fusion rule to the target critics $\{Z_{\bar{\phi}_i}(s', a')\}_{i=1}^{N}$ yields (μ'_F, v'_F). Let $\gamma_d = \gamma(1 - d)$. The Gaussian Bellman target is:

$$Z_{\text{tar}} = \mathcal{N}(\mu_{\text{tar}}, v_{\text{tar}}) \tag{24}$$

where

$$\mu_{\text{tar}} = r + \gamma_d \mu'_F \tag{25}$$

$$v_{\text{tar}} = \max\left(\gamma_d^2 v'_F,\, v_{\min}\right) \tag{26}$$

The variance floor $v_{\min} > 0$ prevents a degenerate Gaussian target on terminal transitions. Each critic is trained using:

$$\mathcal{L}_{\text{dist},i} = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left[D_{\text{KL}} \left(\text{sg}[Z_{\text{tar}}] \parallel Z_{\phi_i}(s, a) \right) \right] \tag{27}$$

where $\text{sg}[\cdot]$ denotes stop-gradient.

3.5 Stability-Triggered Guidance

STG modifies only the critic weights used for actor guidance. Unless stated otherwise, all quantities in this subsection are evaluated at the current policy action. We normalize the ensemble-relative disagreement score as:

$$\widetilde{\delta}_i(s, a) = \frac{\delta_i(s, a)}{\bar{\delta}(s, a) + \epsilon_\delta} \tag{28}$$

and define disagreement-aware reliability weights by:

$$\beta_i(s, a) = \text{softmax}_i \left(\frac{\log \alpha_i(s, a) - \lambda_\delta \widetilde{\delta}_i(s, a)}{T_\beta} \right) \tag{29}$$

where $\lambda_\delta \geq 0$ controls the disagreement penalty and $T_\beta > 0$ is a temperature. Because $\alpha_i(s, a) > 0$, this parameterization is well defined. When $\lambda_\delta = 0$ and $T_\beta = 1$, it reduces to $\beta_i(s, a) = \alpha_i(s, a)$. The ensemble-level trigger and interpolation coefficient are:

$$h(s, a) = 1 - \exp\left(-\frac{\bar{\delta}(s, a)}{\tau_h} \right) \tag{30}$$

where $\tau_h > 0$, and $\eta(s, a) = \eta_{\max} h(s, a), 0 \leq \eta_{\max} \leq 1$. The actor-guidance weights are:

$$\omega_i(s, a) = (1 - \eta(s, a))\alpha_i(s, a) + \eta(s, a)\beta_i(s, a) \tag{31}$$

Thus ω_i is a normalized convex weighting that remains close to precision fusion under low disagreement and shifts toward β_i as disagreement increases. During actor optimization, α_i, β_i, η, and ω_i are treated as stop-gradient quantities. STG is therefore a semi-gradient actor-guidance rule rather than the exact gradient of an objective that also optimizes the reliability mechanism through the actor.

3.6 Action-Gradient Policy Alignment

STG controls how strongly each critic contributes to the actor update, but it does not directly constrain whether the resulting action gradients point in compatible directions. AGPA addresses this complementary failure mode by reducing excessive directional conflict only when a sufficiently coherent actor-relevant consensus remains identifiable. During critic optimization, AGPA evaluates the critics at the current policy action while preventing the alignment loss from directly updating the actor:

$$a_\theta^{\mathrm{sg}}(s) = \mathrm{sg}\left[\pi_\theta(s)\right] \tag{32}$$

$$g_i(s) = \nabla_a \mu_i(s,a)|_{a=a_\theta^{\mathrm{sg}}(s)} \tag{33}$$

The derivative graph of $g_i(s)$ is retained so that the AGPA loss can differentiate through the action gradient with respect to ϕ_i. In the implementation, the active set $\mathcal{I}(s)$, its size $M(s)$, and the normalized directions $u_i(s)$ are defined as above. If $M(s) < 2$, AGPA is not applied. AGPA measures disagreement using the critic weights actually delivered to the actor. For $i \in \mathcal{I}(s)$, we renormalize these weights over the active critics:

$$\widehat{\omega}_i(s) = \frac{\mathrm{sg}\left[\omega_i\left(s, a_\theta^{\mathrm{sg}}(s)\right)\right]}{\sum_{j\in\mathcal{I}(s)} \mathrm{sg}\left[\omega_j\left(s, a_\theta^{\mathrm{sg}}(s)\right)\right]} \tag{34}$$

Detaching $\widehat{\omega}_i$ prevents the alignment loss from being reduced by changing reliability variables instead of action-gradient directions. The actor-relevant resultant and its corresponding dispersion are:

$$m_\omega(s) = \sum_{i\in\mathcal{I}(s)} \widehat{\omega}_i(s) u_i(s) \tag{35}$$

$$R_\omega(s) = \|m_\omega(s)\|_2 \tag{36}$$

$$A_\omega(s) = \sum_{i\in\mathcal{I}(s)} \widehat{\omega}_i(s) \|u_i(s) - m_\omega(s)\|_2^2 = 1 - R_\omega(s)^2 \tag{37}$$

AGPA preserves low-to-moderate disagreement through a dead-zone threshold A_0 and abstains when the weighted directions nearly cancel. For states with $M(s) \geq 2$, define:

$$D_A(s) = \max\left\{A_\omega(s) - A_0,\ 0\right\} \tag{38}$$

and define the gate as:

$$\Gamma(s) = \mathrm{sg}\left[\left(1 - \exp\left\{-\kappa_A D_A(s)\right\}\right) \mathbf{1}\left\{R_\omega(s) \geq r_0\right\}\right] \tag{39}$$

where $\kappa_A > 0$ controls the gate sensitivity and $0 < r_0 < 1$ specifies the minimum consensus strength. For $M(s) < 2$, we set $\Gamma(s) = 0$. Since $A_\omega(s) = 1 - R_\omega(s)^2$, a nonzero intervention requires:

$$A_0 < A_\omega(s) \leq 1 - r_0^2 \tag{40}$$

Accordingly, the thresholds are chosen to satisfy:

$$0 \leq A_0 < 1 - r_0^2 \tag{41}$$

The lower boundary preserves ordinary ensemble diversity, whereas the upper boundary prevents AGPA from forcing alignment when no sufficiently coherent consensus can be identified. Within this gated region, AGPA promotes directional concentration through:

$$\mathcal{L}_{\mathrm{AGPA}} = \mathbb{E}_{s\sim\mathcal{D}}\left[\Gamma(s)\left(1 - R_\omega(s)\right)\right] \tag{42}$$

The complete critic objective is:

$$\mathcal{L}_{\mathrm{critic}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{dist},i} + \lambda_A \mathcal{L}_{\mathrm{AGPA}} \tag{43}$$

where $\lambda_A \geq 0$ controls the strength of the alignment regularization. STG therefore controls critic contributions to actor guidance, whereas AGPA regularizes the geometry of the gradients that remain actor-relevant. In particular, AGPA may activate even when value-level disagreement is small, allowing DAGA to directly address value–gradient mismatch without enforcing unconditional gradient homogenization.

3.7 Actor Update and Training Procedure

The actor maximizes the detached weighted critic objective:

$$\mathcal{J}_{\text{actor}}(\theta) = \mathbb{E}_{s\sim\mathcal{D}}\left[\sum_{i=1}^{N} \text{sg}\left[\omega_i(s,\pi_\theta(s))\right]\mu_i(s,\pi_\theta(s))\right] \tag{44}$$

Holding the critics and guidance weights fixed gives:

$$\nabla_\theta \mathcal{J}_{\text{actor}} = \mathbb{E}_{s\sim\mathcal{D}}\left[J_\theta(s)^\top \sum_{i=1}^{N} \text{sg}[\omega_i(s)] g_i(s)\right] \tag{45}$$

where $\omega_i(s)$ abbreviates evaluation at $a = \pi_\theta(s)$. Actor updates are delayed as in the underlying TD3-style backbone, and target networks are updated by Polyak averaging. Each training iteration therefore consists of Bellman-target construction, distributional critic optimization with gated AGPA, delayed actor optimization with STG weights, and target-network updates. Local properties of value–gradient mismatch, reliability weighting, STG, and AGPA are established in Appendix A.

4 Experiments

4.1 Experimental Setup

We conduct detailed mechanism, ablation, and sensitivity analyses on h1hand-hurdle-v0, a high-dimensional whole-body control task with strongly nonlinear dynamics. This task serves as the primary diagnostic setting for examining value–gradient mismatch, critic action-gradient geometry, and the interaction between STG and AGPA, and we evaluate cross-task performance on T1JoystickRoughTerrain, LeapCubeReorient, h1hand-reach-v0, h1hand-basketball-v0, Go1Getup, and CheetahRun (Sferrazza et al., 2024; Todorov et al., 2012).

Baselines include PPO (Schulman et al., 2017), SAC (Haarnoja et al., 2018), DDPG (Lillicrap et al., 2015), TD3 (Fujimoto et al., 2018), REDQ (Chen et al., 2021), and DAGA's precision-fusion Backbone. We also compare against the recent continuous-control baseline SimbaV2 (Lee et al., 2025). The focused ablation compares Backbone, Backbone+STG, Backbone+AGPA, and DAGA.

Main comparisons use ten seeds. Each checkpoint is evaluated over ten episodes every 5,000 environment steps. Learning curves report the mean across seeds with approximate 95% confidence intervals based on the standard error of the mean (SEM), while Table 1 reports mean ± standard deviation. Returns are reported separately because reward scales differ across tasks. Complete DAGA-specific and training hyperparameters are reported in Tables 2 and 3.

4.2 Cross-Task Performance

Figure 2 compares DAGA with its Backbone and strong baselines, while Table 1 additionally includes PPO, SAC, and DDPG. DAGA attains the highest mean final return on four of six tasks and achieves a higher mean final return than its Backbone on all six. The largest absolute gains over the Backbone occur on H1-Reach and Cheetah-Run. On T1-Rough and H1-Basket, PPO and SimbaV2 respectively achieve the highest mean final returns.

4.3 Mechanism and Component Ablation

Figure 3 examines h1hand-hurdle-v0 using the diagnostics defined in Section 3.3. Diagnostic quantities are computed on the same replay-state batch at each checkpoint. Figure 3(c) bins states by within-run $\bar{\delta}$ quantiles, non-perfect C_{mean} and nonzero R_{conf} in the lowest-disagreement bin show that distributional agreement does not imply action-gradient agreement. Figure 3(a) compares Backbone, Backbone+STG, Backbone+AGPA, and DAGA.

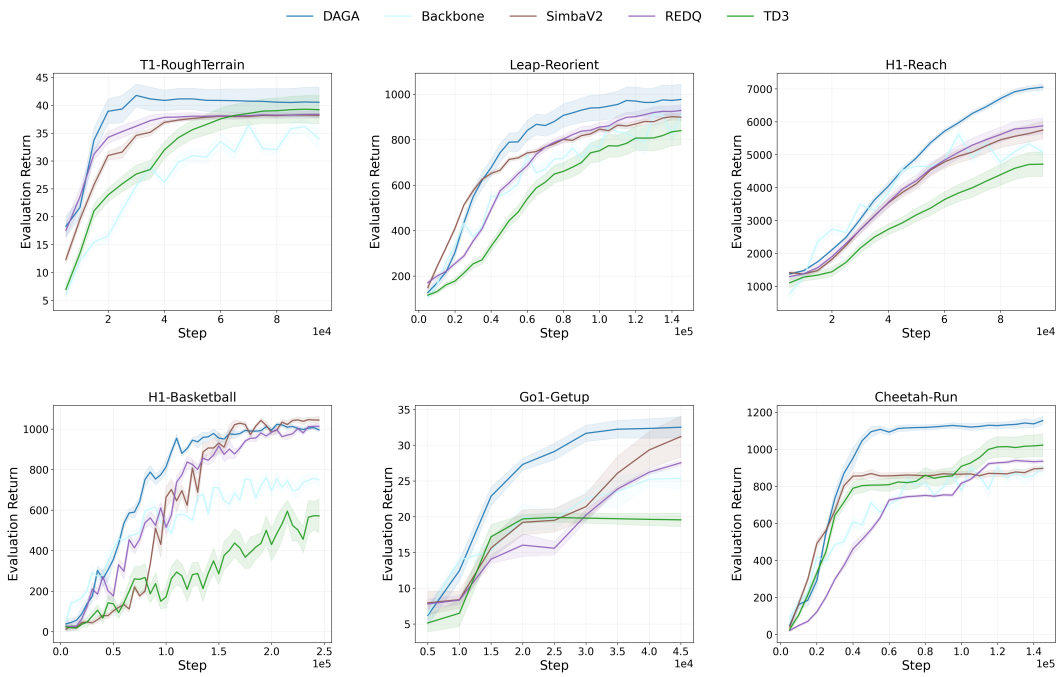

Figure 2: Evaluation return across six continuous-control tasks over ten independent seeds. Curves show seed means, with shaded regions indicating approximate 95% confidence intervals based on SEM.

Table 1: Final evaluation returns mean ± standard deviation over ten independent seeds. Boldface indicates the highest mean return in each task. Task names are abbreviated for compactness.

Method	T1-Rough	Leap-Reorient	H1-Reach	H1-Basket	Go1-Getup	Cheetah-Run
PPO	**43.2** ± 5.7	743.8 ± 110.4	4955.5 ± 354.6	822.3 ± 17.6	20.1 ± 2.6	838.6 ± 116.8
SAC	39.4 ± 1.3	377.1 ± 18.3	4762.0 ± 680.7	915.4 ± 78.6	18.8 ± 0.8	758.1 ± 36.3
DDPG	30.9 ± 0.7	272.6 ± 65.4	4196.6 ± 171.1	585.6 ± 120.5	19.0 ± 1.1	795.3 ± 15.1
TD3	39.3 ± 4.0	830.9 ± 97.8	4707.7 ± 599.6	533.3 ± 108.9	19.6 ± 1.5	1019.0 ± 96.5
REDQ	38.4 ± 1.0	925.6 ± 39.5	5848.4 ± 385.9	1004.4 ± 17.9	27.5 ± 0.7	934.9 ± 13.4
SimbaV2	38.2 ± 0.5	898.4 ± 23.4	5695.6 ± 425.2	**1043.7** ± 16.1	31.2 ± 4.6	888.4 ± 23.0
Backbone	35.1 ± 2.4	879.0 ± 52.6	5211.1 ± 82.1	738.1 ± 25.6	25.4 ± 2.4	870.5 ± 32.2
DAGA	40.6 ± 4.3	**974.9** ± 108.0	**7029.1** ± 143.7	1001.7 ± 13.2	**32.5** ± 2.4	**1144.7** ± 36.7

Both component variants achieve higher mean returns than the Backbone on the primary task, while full DAGA achieves the highest mean return among the four ablation variants. Figure 3(b) reports actor-relevant R_w ($w = \alpha$ without STG and $w = \omega$ with STG) together with G_{norm}. AGPA-based variants exhibit greater actor-relevant directional concentration while maintaining comparable gradient magnitudes, consistent with the intended mechanism.

4.4 Sensitivity Analysis

Figure 4 reports five-seed sensitivity experiments on h1hand-hurdle-v0. Joint variation of $\eta_{\max}$ and λ_A shows stronger performance at moderate settings than when either mechanism is disabled or overly strong. Ensemble-size performance is non-monotonic: moderate N performs well, whereas $N = 10$ provides no additional benefit under the tested configuration. Because changing N also changes capacity and cost, we interpret this only as ensemble-size sensitivity.

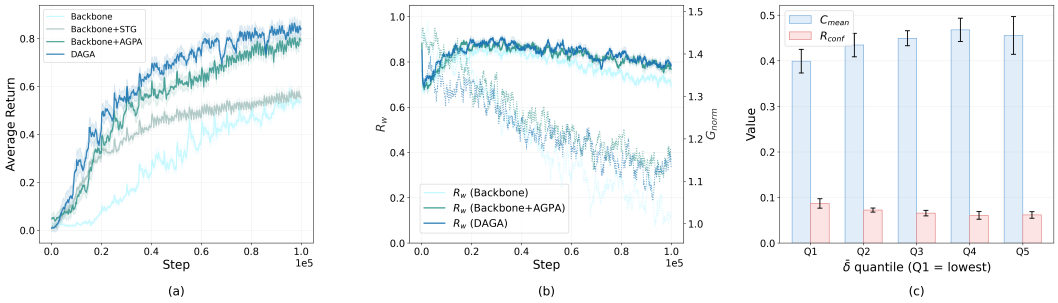

Figure 3: Mechanism analysis on h1hand-hurdle-v0. (a) Component ablation. (b) Actor-relevant resultant length R_w (solid, left axis) and gradient magnitude G_{norm} (dotted, right axis). (c) Pairwise cosine similarity C_{mean} and conflict ratio R_{conf} across δ quantiles with Q1 denoting the lowest-disagreement states.

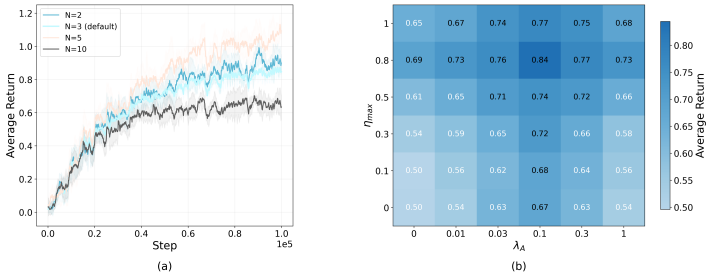

Figure 4: Sensitivity on h1hand-hurdle-v0. (a) Learning curves for critic-ensemble size N. (b) Final mean return under joint variations of STG strength $\eta_{\max}$ and AGPA strength λ_A. The zero settings correspond to disabling the respective intervention or regularization, while intermediate settings generally yield stronger performance than either very weak or overly strong configurations.

5 Conclusion

We introduced DAGA to address value–gradient mismatch in multi-critic deterministic actor–critic learning. STG adapts actor guidance using ensemble-relative distributional reliability, while AGPA selectively regularizes actor-relevant gradient geometry when a coherent consensus remains. Primary-task ablations support the complementary roles of STG and AGPA, and full DAGA achieves higher mean final returns than its Gaussian multi-critic Backbone across all six cross-task evaluations. Together with the local analysis, these results suggest that action-gradient geometry can complement value-level ensemble aggregation.

Several limitations remain. DAGA does not guarantee that critic consensus is correct, AGPA may therefore reinforce a coherent but biased direction. STG uses ensemble-relative distributional disagreement as a guidance signal rather than a calibrated estimate of gradient quality, and its benefit may depend on critic diversity. Component-level analysis is concentrated on one high-dimensional control task, while cross-task experiments evaluate the full method. Finally, increasing ensemble size does not monotonically improve performance and also increases computational cost.

AI Use Statement

Generative AI tools were used to assist with literature search, manuscript editing and organization, and interpretation and presentation of experimental results. All AI-assisted content, including mathematical statements, technical descriptions, citations, and experimental interpretations, was reviewed and verified by the authors. The authors take full responsibility for the correctness and content of the submission.

References

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble, volume 34. 2021.

Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. 2018.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. 2017.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. arXiv preprint arXiv:2101.05982, 2021.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression, volume 32. 2018.

Pierluca D'Oro and Wojciech Jaśkowski. How to learn a useful critic? model-based action-gradient-estimator policy optimization, volume 33. 2020.

Linjiajie Fang, Ruoxue Liu, Jing Zhang, Wenjia Wang, and Bingyi Jing. Diffusion actor-critic: Formulating constrained policy iteration as diffusion noise regression for offline reinforcement learning, volume 2025. 2025.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In International conference on machine learning, pp. 1587–1596. 2018.

Swaminathan Gurumurthy, Zachary Manchester, and J Zico Kolter. Practical critic gradient based actor critic for on-policy reinforcement learning. 2023.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905, 2018.

Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. 2020.

Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. Hyperspherical normalization for scalable deep reinforcement learning. arXiv preprint arXiv:2502.15280, 2025.

Jeong Woon Lee, Kyoleen Kwak, Daeho Kim, and Hyoseok Hwang. Stabilizing the Q-Gradient Field for Policy Smoothness in Actor-Critic Methods. 2026.

Qiyang Li and Sergey Levine. Q-learning with adjoint matching. 2026.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning, volume 34. 2021.

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

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. arXiv preprint arXiv:2403.10506, 2024.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

David Valencia, Henry Williams, Yuning Xing, Trevor Gee, Bruce A MacDonald, and Minas Liarokapis. CTD4–A deep continuous distributional actor-critic agent with a kalman fusion of multiple critics, volume 39. 2025.

Nicklas Werge, Yi-Shan Wu, Bahareh Tasdighi, and Melih Kandemir. Directional ensemble aggregation for actor-critics. 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, volume 33. 2020.

Yiwen Zhu, Jinyi Liu, Wenya Wei, Qianyi Fu, Yujing Hu, Zhou Fang, Bo An, Jianye Hao, Tangjie Lv, and Changjie Fan. vmfer: von mises-fisher experience resampling based on uncertainty of gradient directions for policy improvement of actor-critic algorithms. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 2621–2623, 2024.

A Appendix

The following results characterize local properties of the diagnostics and update rules used by DAGA. They do not establish global convergence or performance guarantees for nonlinear deep actor-critic learning.

A.1 Pointwise Value Agreement Does Not Imply Gradient Agreement

Proposition 1 (Pointwise decoupling). Let $a_0 \in \mathbb{R}^{d_a}$. For arbitrary nonzero vectors $q_1, q_2 \in \mathbb{R}^{d_a}$ and scalar c, define:

$$\mu_1(a) = c + q_1^\top (a - a_0), \qquad \mu_2(a) = c + q_2^\top (a - a_0) \tag{46}$$

Then

$$\mu_1(a_0) = \mu_2(a_0) = c, \tag{47}$$

while

$$\nabla_a \mu_1(a_0) = q_1, \qquad \nabla_a \mu_2(a_0) = q_2. \tag{48}$$

Hence, pointwise value agreement places no restriction on the two local action gradients. In particular, choosing $q_2 = -q_1$ yields exact value agreement with opposing action-gradient directions.

Proposition 1 establishes the formal possibility of value–gradient mismatch but does not imply that large mismatch is ubiquitous or necessarily harmful. The conclusion changes when value agreement holds over a neighborhood.

Proposition 2 (Local value agreement constrains gradient disagreement). Fix a state and let

$$h(a) = \mu_i(a) - \mu_j(a) \tag{49}$$

Suppose that h has an L-Lipschitz continuous gradient on $B(a_0, r)$, with $L > 0$, and satisfies:

$$|h(a)| \le \varepsilon, \qquad \forall a \in B(a_0, r). \tag{50}$$

Then, for every $t \in (0, r]$,

$$\|\nabla h(a_0)\|_2 \le \frac{2\varepsilon}{t} + \frac{Lt}{2}. \tag{51}$$

If $\varepsilon > 0$ and $r \ge 2\sqrt{\frac{\varepsilon}{L}}$, choosing $t = 2\sqrt{\varepsilon/L}$ gives:

$$\|\nabla \mu_i(a_0) - \nabla \mu_j(a_0)\|_2 \le 2\sqrt{L\varepsilon}. \tag{52}$$

Proof. If $\nabla h(a_0) = 0$, the result is immediate. Otherwise, let $q = \nabla h(a_0)/\|\nabla h(a_0)\|_2$. By L-smoothness

$$h(a_0 - tq) \le h(a_0) - t\|\nabla h(a_0)\|_2 + \frac{L}{2}t^2 \tag{53}$$

Using $h(a_0 - tq) \ge -\varepsilon$ and $h(a_0) \le \varepsilon$ gives:

$$\|\nabla h(a_0)\|_2 \le \frac{2\varepsilon}{t} + \frac{Lt}{2} \tag{54}$$

Minimizing the right-hand side over t gives $t = 2\sqrt{\varepsilon/L}$ whenever this choice lies in $(0, r]$, which yields the stated special case. □

Propositions 1–2 clarify the scope of value–gradient mismatch: pointwise value agreement alone does not constrain critic action gradients, whereas sufficiently strong local functional agreement together with smoothness does.

A.2 Bias-Variance Characterization of Reliability Weighting

At a fixed state–action pair, let:

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)^\top = Q^\pi(s,a)\mathbf{1} + \mathbf{b} + \mathbf{e} \tag{55}$$

where $\mathbf{b}, \mathbf{e} \in \mathbb{R}^N$, $\mathbb{E}[\mathbf{e}] = 0$, and $\mathrm{Cov}(\mathbf{e}) = \Sigma_e$. For deterministic weights w satisfying $\mathbf{1}^\top w = 1$, define:

$$\widehat{Q}_w = w^\top \boldsymbol{\mu}. \tag{56}$$

Then

$$\mathrm{Bias}\left(\widehat{Q}_w\right) = w^\top \mathbf{b} \tag{57}$$

$$\mathrm{Var}\left(\widehat{Q}_w\right) = w^\top \Sigma_e w \tag{58}$$

and

$$\mathrm{MSE}\left(\widehat{Q}_w\right) = \left(w^\top \mathbf{b}\right)^2 + w^\top \Sigma_e w \tag{59}$$

Thus, reliability reweighting can alter both bias and variance. Without assumptions relating the reliability scores to $\mathbf{b}$ and Σ_e, no weighting rule universally dominates another. If the critics are unbiased and have independent estimation errors:

$$\Sigma_e = \mathrm{diag}\left(\sigma_1^2, \dots, \sigma_N^2\right) \tag{60}$$

the minimum-variance unbiased weights are:

$$w_i^\star = \frac{\sigma_i^{-2}}{\sum_j \sigma_j^{-2}} \tag{61}$$

More generally, if Σ_e is positive definite and no additional nonnegativity constraint is imposed:

$$w^\star = \frac{\Sigma_e^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma_e^{-1}\mathbf{1}} \tag{62}$$

Ignoring the numerical stabilizer, the precision weights $\alpha_i \propto v_i^{-1}$ recover the independent-error solution only if the predicted return variances are proportional to calibrated estimation-error variances. Shared replay data and bootstrap targets generally induce correlated errors, and return-distribution variance need not equal estimation-error variance. Precision fusion is therefore a conditional statistical motivation rather than an unconditional optimality result. The same decomposition applies locally to the frozen STG mixture:

$$\omega = (1 - \eta)\alpha + \eta\beta \tag{63}$$

Because practical STG weights depend on critic outputs, this is a frozen-weight characterization rather than a complete statistical model of the adaptive weighting rule.

A.3 Bounded Intervention of STG

For notational brevity, let all weights below be evaluated at the current policy action and define:

$$g_\alpha(s) = \sum_i \alpha_i(s) g_i(s) \tag{64}$$

$$g_\beta(s) = \sum_i \beta_i(s) g_i(s) \tag{65}$$

$$g_\omega(s) = (1 - \eta(s)) g_\alpha(s) + \eta(s) g_\beta(s) \tag{66}$$

Then

$$\|g_\omega(s) - g_\alpha(s)\|_2 = \eta_{\max} \left(1 - e^{-\bar{\delta}(s)/\tau_h}\right) \|g_\beta(s) - g_\alpha(s)\|_2 \tag{67}$$

and, using $1 - e^{-x} \le x$ for $x \ge 0$,

$$\|g_\omega(s) - g_\alpha(s)\|_2 \le \eta_{\max} \frac{\bar{\delta}(s)}{\tau_h} \|g_\beta(s) - g_\alpha(s)\|_2 \tag{68}$$

Thus, STG vanishes when ensemble-relative disagreement vanishes and provides a bounded interpolation away from precision-fused actor guidance. Let:

$$p_\alpha(s) = J_\theta(s)^\top g_\alpha(s) \tag{69}$$

$$p_\omega(s) = J_\theta(s)^\top g_\omega(s) \tag{70}$$

Then

$$\|p_\omega(s) - p_\alpha(s)\|_2 \le \|J_\theta(s)\|_2 \|g_\omega(s) - g_\alpha(s)\|_2 \tag{71}$$

where $\|J_\theta(s)\|_2$ denotes the induced operator norm. Hence action-space disagreement need not be preserved by the policy Jacobian, but the resulting change in the actor-parameter update is controlled by its local operator norm.

A.4 Geometry and First-Order Effect of AGPA

Fix a state at which the AGPA gate is active. Then $R_\omega \ge r_0 > 0$, and the detached coefficients $\widehat{\omega}_i$ may be treated as constants. Define:

$$m_\omega = \sum_{i \in \mathcal{I}(s)} \widehat{\omega}_i u_i \tag{72}$$

$$R_\omega = \|m_\omega\|_2 \tag{73}$$

$$d_\omega = \frac{m_\omega}{R_\omega} \tag{74}$$

The weighted dispersion and alignment identities are:

$$A_\omega = 1 - R_\omega^2 \tag{75}$$

$$1 - R_\omega = \sum_{i \in \mathcal{I}(s)} \widehat{\omega}_i \left(1 - u_i^\top d_\omega\right) \tag{76}$$

Thus, resultant length, angular dispersion, and the weighted cosine-alignment objective describe the same actor-relevant directional concentration. Moreover, an active gate requires:

$$A_0 < A_\omega \le 1 - r_0^2 \tag{77}$$

which gives AGPA both a lower dead zone and an upper no-consensus region. Let

$$\ell_{\text{geo}} = 1 - R_\omega \tag{78}$$

For $i \in \mathcal{I}(s)$,

$$\nabla_{g_i} \ell_{\text{geo}} = -\frac{\widehat{\omega}_i}{\|g_i\|_2} \left(I - u_i u_i^\top\right) d_\omega. \tag{79}$$

Therefore,

$$g_i^\top \nabla_{g_i} \ell_{\text{geo}} = 0 \tag{80}$$

The direct AGPA gradient in action-gradient space is thus orthogonal to the current g_i and has no first-order radial component. Multiplication by the detached gate $\Gamma(s)$ preserves this property. For a hypothetical direct gradient-space step, define

$$g_i^{\text{new}} = g_i - \rho \nabla_{g_i} \ell_{\text{geo}} \tag{81}$$

Then

$$\|g_i^{\text{new}}\|_2^2 = \|g_i\|_2^2 + \rho^2 \|\nabla_{g_i} \ell_{\text{geo}}\|_2^2 \tag{82}$$

and consequently

$$\frac{d}{d\rho} \|g_i^{\text{new}}\|_2 \Big|_{\rho=0} = 0 \tag{83}$$

Hence the direct first-order effect of AGPA is directional rather than radial. This is a local property in action-gradient space, not a claim that neural-network critic updates preserve $\|g_i\|_2$: Bellman optimization, parameterization, optimizer preconditioning, finite steps, and changing gates can all affect gradient magnitudes.

Table 2: DAGA-specific hyperparameters and numerical constants. The main sensitivity analysis jointly varies $\eta_{\max}$ and λ_A.

Group	Symbol	Role	Default
STG	λ_δ	STG disagreement scale	0.5
	T_β	STG temperature	1.0
	τ_h	STG activation threshold	0.1
	$\eta_{\max}$	Maximum STG intervention	0.8
AGPA	A_0	Angular-dispersion threshold	0.2
	r_0	Consensus threshold	0.8
	κ_A	AGPA gate sensitivity	10
	λ_A	AGPA regularization strength	0.1
Numerical	ϵ_v	Variance stabilizer	10^{-6}
	ϵ_δ	Disagreement stabilizer	10^{-6}
	ϵ_g	Gradient-norm threshold	10^{-6}
	$v_{\min}$	Minimum predictive variance	10^{-6}

Table 3: Training and architecture hyperparameters used for DAGA. Unless otherwise specified, the same configuration is used across tasks.

Hyperparameter	Value
Batch size	8192
Final actor learning rate	3×10^{-5}
Final critic learning rate	3×10^{-5}
Weight decay	0
Critic hidden dimension	1024
Critic hidden blocks	2
Actor hidden dimension	256
Actor hidden blocks	1
Critic dropout rate	0.1
Policy update frequency	2
Target update coefficient τ	0.1
Gradient updates per iteration	2
Default ensemble size N	3
Evaluation interval	5,000
Evaluation episodes	10
Main-experiment seeds	10
Sensitivity-analysis seeds	5
Discount factor γ	0.99
Replay-buffer capacity	10^6
Target-policy noise std σ_{tps}	0.2
Target-policy noise clip	0.5
Optimizer	Adam