CAFRL: CAUSALITY-AWARE FLOW BASED POLICY FOR ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-based policies enhance maximum-entropy reinforcement learning by representing complex and multimodal behaviors through learned transport dynamics. However, entropy evaluation of implicit flow policies is generally difficult, motivating path-space regularization as an alternative to explicit density-based constraints. Existing approaches typically adopt isotropic exploration geometries, assigning identical kinetic costs to different action dimensions despite their heterogeneous relevance to tasks.We introduce CAFRL, a causality-aware regularization framework that incorporates action-to-reward effects into the exploration geometry of flow policies. CAFRL derives action relevance from interaction data through causal analysis and maps the resulting effects into a reference process, inducing anisotropic path kinetic regularization. This enables task-relevant directions to deviate from the reference process with adaptive path costs while maintaining a consistent regularization scale, thereby inducing a causal anisotropic exploration geometry. Across eight HumanoidBench tasks with five random seeds for each method, CAFRL achieves the highest average return compared with FLAC, ACE, and SimbaV2 after one million environment steps. CAFRL improves over the strongest baseline on each task by a median margin of 80.3%. These results support the empirical performance of the CAFRL configuration evaluated.

1 INTRODUCTION

Maximum-entropy reinforcement learning jointly optimizes expected return and policy entropy, providing a principled framework for learning stochastic policies that balance task performance and behavioral diversity. Soft Q-learning and Soft Actor–Critic (SAC) have demonstrated the effectiveness of this paradigm in continuous-action settings (Haarnoja et al., 2017; 2018). More recently, flow-based policies have expanded the expressiveness of stochastic policies by learning continuous transformations that transport simple initial distributions into complex and multimodal action distributions (Lipman et al., 2023; McAllister et al., 2026). However, exploiting the flexibility of flow-based policies while maintaining effective and task-dependent exploration remains challenging.

A key difficulty is that flow policies generate actions through continuous transformations, while their terminal action densities are generally unavailable directly from forward sampling. Consequently, entropy regularization based on explicit log densities becomes difficult to apply. Recent approaches have investigated alternative formulations for optimizing flow-based policies, including policy gradients without explicit likelihood computation and continuous normalizing flow based policy optimization (McAllister et al., 2026; Yang et al., 2026). However, these approaches mainly focus on improving optimization tractability and do not explicitly consider how exploration should adapt to the task relevance of different action dimensions.

In high-dimensional continuous control, different action dimensions often have unequal effects on future rewards. Applying an identical exploration geometry to all action dimensions may therefore introduce unnecessary exploration in irrelevant directions while limiting exploration along task-critical dimensions. Recent studies have shown that causal information can provide action-relevant signals for improving reinforcement learning efficiency (Ji et al., 2024; Seitzer et al., 2021), while flow-based exploration methods have demonstrated the potential of learned continuous action-generation processes for high-dimensional control (Wei et al., 2026). This motivates a fundamental

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.

question: *how can causal action relevance be incorporated into the geometry of flow-based action generation to achieve task-dependent exploration without requiring explicit action densities?*

A challenge in addressing this problem is that causal relevance cannot be obtained from statistical correlation alone and requires estimating structural relationships from interaction data. Methods such as DirectLiNGAM provide causal effect estimation under linear non-Gaussian structural assumptions (Shimizu et al., 2011). Furthermore, when modifying stochastic exploration processes, causal weighting should be introduced consistently into both the reference process and the corresponding path-space cost, rather than being applied as an external heuristic. Schrödinger-bridge formulations provide a principled connection between reference stochastic processes, path-space divergence, and kinetic energy regularization (Léonard, 2014; Liu et al., 2024).

We introduce CAFRL, a causality-aware framework for flow-based reinforcement learning that incorporates estimated causal effects into the geometry of policy generation. CAFRL constructs a causality-aware reference diffusion process for flow-matching action generation, where action dimensions with higher estimated task relevance receive larger adaptive exploration scales and reduced relative path costs. This formulation enables the policy to allocate exploration according to causal task structure while preserving the stochastic flow-based action-generation framework.

Our contributions are threefold.

- We propose a causality-aware exploration mechanism for flow-based reinforcement learning, where estimated causal effects characterize the task relevance of individual action dimensions and provide adaptive guidance for stochastic exploration.
- We introduce a causal reference diffusion formulation for flow-matching policy generation, allowing causal action relevance to modify the exploration geometry through a principled path-space formulation rather than heuristic action weighting.
- We analyze the proposed causal action-generation framework and evaluate its effectiveness on eight HumanoidBench tasks, demonstrating improved performance over existing flow-based and causal reinforcement learning baselines.

2 Related Work

Maximum-entropy reinforcement learning and flow-based policies. Maximum-entropy reinforcement learning augments expected return with policy entropy, improving exploration and robustness in continuous-action problems. Soft Actor–Critic (SAC) is a representative approach that optimizes this objective with tractable stochastic policies (Haarnoja et al., 2018). Flow matching provides a more expressive alternative by transporting a simple base distribution to a complex action distribution through learned continuous dynamics (Lipman et al., 2023). However, the terminal density of an implicit flow policy is generally not available from forward sampling alone, which complicates the direct use of Shannon-entropy regularization. FLAC addresses this issue by replacing explicit terminal-density regularization with a path-space kinetic-energy objective (Lv et al., 2026).

Causal information for reinforcement learning. Causal reinforcement learning has explored how structural information can improve representation learning, exploration, and sample efficiency by identifying variables that are controllable or relevant to reward (Seitzer et al., 2021). A related line of work estimates action–reward dependencies from interaction data and uses them to characterize the relative importance of action dimensions. DirectLiNGAM provides one such estimator under a linear acyclic structural model with independent non-Gaussian disturbances (Shimizu et al., 2011). These methods provide estimates of causal structure or effect magnitude, but do not by themselves determine how such information should alter the stochastic geometry of an implicit policy.

3 Preliminaries

3.1 Maximum-Entropy Reinforcement Learning and Flow Policies

Consider a continuous-action Markov decision process (MDP), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively. Maximum-entropy reinforcement learning augments

cumulative reward with a policy-entropy term to encourage exploration:

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)+\alpha\mathcal{H}(\pi(\cdot|s_t))\right)\right], \tag{1}$$

Here, $\mathcal{H}(\cdot)$ denotes policy entropy and α is the temperature coefficient.

Flow-matching policies represent complex action distributions through continuous-time dynamical systems. Given a state s, a flow policy uses a state-dependent velocity field to transform a simple reference distribution into an action distribution:

$$\frac{dX_\tau}{d\tau} = v_\theta(X_\tau, s, \tau), \qquad A = T(X_1), \tag{2}$$

Here, v_θ is a learnable velocity field and $T(\cdot)$ maps the terminal flow variable to an action.

FLAC interprets policy generation as a path-space optimization problem and uses path kinetic energy to regularize deviation from a reference process:

$$\mathcal{E}_{0,\theta}(s) = \frac{1}{2}\mathbb{E}\left[\int_0^1 \|v_\theta(X_\tau, s, \tau)\|_2^2 d\tau\right]. \tag{3}$$

This energy corresponds to an isotropic Euclidean metric in action space, assigning the same path-change cost to different action directions. FLAC thus provides a maximum-entropy policy-optimization framework based on path-space regularization.

3.2 Causal Structural Model

Following the causal policy–reward structural formulation of ACE (Ji et al., 2024), we consider interaction tuples (s_t, a_t, r_t), where $s_t \in \mathcal{S}$, $a_t \in \mathcal{A} \subseteq \mathbb{R}^{d_A}$, and $r_t \in \mathbb{R}$ denote the state, action, and reward, respectively. The reward-generating mechanism can be represented as

$$r_t = r_{\mathcal{M}}\left(\boldsymbol{B}_{s\to r|a} \odot s_t,\ \boldsymbol{B}_{a\to r|s} \odot a_t,\ \epsilon_t\right), \tag{4}$$

where $\boldsymbol{B}_{s\to r|a}$ and $\boldsymbol{B}_{a\to r|s}$ describe the structural relations from state and action variables to the reward, $\odot$ denotes element-wise multiplication, and ϵ_t represents exogenous noise.

This formulation allows different primitive action dimensions to have different causal influences on task reward. We denote the resulting dimension-wise causal-effect profile by

$$\boldsymbol{c} = (c_1, \ldots, c_{d_A})^\top, \qquad c_i \geq 0, \tag{5}$$

where c_i characterizes the estimated causal relevance of action dimension i to the task objective. The specific estimator used to obtain $\boldsymbol{c}$ from replay data is introduced in Sec. 4.

Under the causal Markov and faithfulness assumptions, together with the absence of unobserved confounders between the modeled variables, the corresponding structural relations can be identified from observational interaction data under the conditions established in prior causal-RL work (Ji et al., 2024). We adopt these assumptions as the causal interpretation underlying the estimated effect profile.

3.3 Generalized Schrödinger Bridge Regularization

A generalized Schrödinger bridge views policy generation as a controlled stochastic process regularized relative to a reference path measure. Let $\mathbb{P}^{\mathrm{ref}}$ denote a reference path measure with initial marginal μ_0, and let $\mathbb{P}$ denote a controlled path measure. Following the formulation adopted in FLAC (Lv et al., 2026), the one-ended generalized Schrödinger bridge is

$$\inf_{\mathbb{P}:\mathbb{P}_0=\mu_0}\left\{\mathbb{E}_{\mathbb{P}}\left[\mathcal{G}_s(X_1)\right] + \lambda D_{\mathrm{KL}}\left(\mathbb{P}\|\mathbb{P}^{\mathrm{ref}}\right)\right\}, \tag{6}$$

where $\mathcal{G}_s$ is a state-conditioned terminal potential and $\lambda > 0$ controls the relative-entropy regularization.

For reinforcement learning, taking

$$\mathcal{G}_s(x) = -Q(s, T(x)) \tag{7}$$

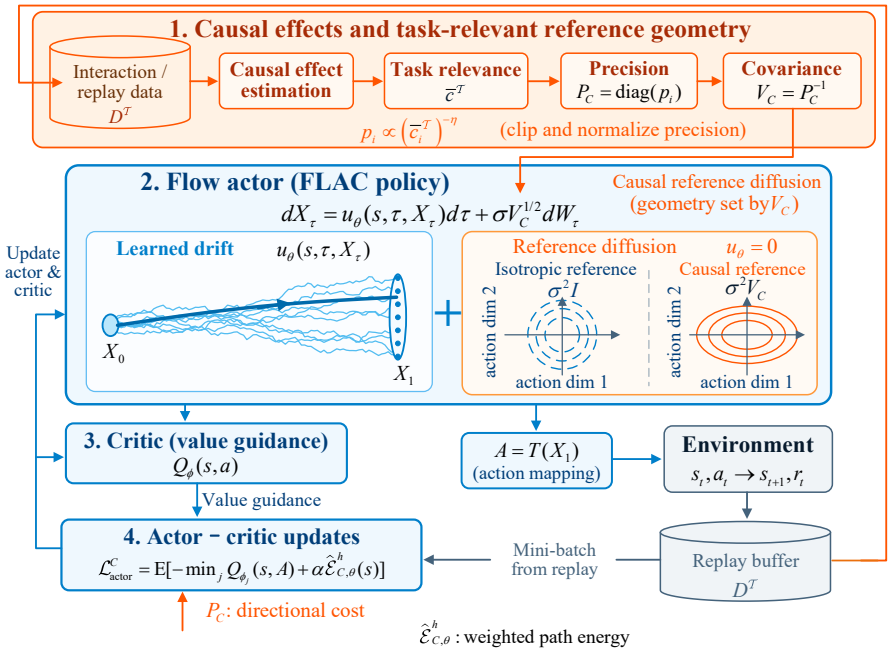


Figure 1: Overview of CAFRL. Causal effects shape the reference geometry of the flow actor, inducing anisotropic exploration, while the SAC critic guides task-oriented policy learning.

encourages the terminal process toward high-value actions while retaining proximity to the reference stochastic process. The corresponding optimal terminal marginal admits the Gibbs form

$$q^\star(x|s) = \frac{1}{Z(s)} \mu_1^{\mathrm{ref}}(x) \exp\left(\frac{Q(s,T(x))}{\lambda}\right), \tag{8}$$

where μ_1^{ref} denotes the terminal marginal of the reference process and $Z(s)$ is the normalizing constant. This standard result follows from the variational characterization of the generalized Schrödinger bridge; see Lv et al. (2026) for the derivation.

Equation (8) highlights the role of the reference process in policy generation: the value function determines the preference for high-value terminal actions, whereas μ_1^{ref} determines the reference geometry against which the policy process is regularized. Standard formulations typically adopt a task-independent isotropic reference process. In the following section, we investigate how causal action relevance can instead be used to construct a task-dependent anisotropic reference geometry.

4 Causal-Aware Flow Reinforcement Learning

CAFRL is a causality-aware flow-based reinforcement learning framework built on an off-policy actor–critic backbone. As illustrated in Fig. 1, it replaces the Gaussian SAC actor with a flow-based stochastic actor while retaining the critic for value guidance.

Interaction data are used to estimate a dimension-wise causal-effect profile $\boldsymbol{c}$, which is mapped to a task-dependent reference geometry with precision matrix P_C and covariance $V_C = P_C^{-1}$. The resulting geometry shapes the reference diffusion of the flow actor and the corresponding directional kinetic cost. The learned drift then transports this process toward high-value actions guided by the SAC critic.

The generated action is executed in the environment and new transitions are collected for iterative actor–critic updates. Thus, causal effects shape how the flow policy explores, while the critic determines which actions are preferred.

4.1 CAUSALITY-AWARE FLOW ACTOR VIA ANISOTROPIC SCHRÖDINGER BRIDGE

Standard SAC typically adopts Gaussian policies, which are convenient but limited in representing complex multimodal action distributions. CAFRL replaces the Gaussian actor with a flow-based stochastic actor and uses causal effects to construct a task-dependent reference geometry.

Let $\bar{\boldsymbol{c}} = (\bar{c}_1, \ldots, \bar{c}_{d_A})^\top$ denote the estimated causal-effect profile. We convert it into dimension-wise precision coefficients

$$p_i = \mathrm{clip}\left(b_C \bar{c}_i^{-\eta}, p_{\min}, p_{\max}\right), \qquad \frac{1}{d_A}\sum_i p_i = 1, \tag{9}$$

and define

$$P_C = \mathrm{diag}(p_1, \ldots, p_{d_A}), \qquad V_C = P_C^{-1}. \tag{10}$$

Therefore, larger causal effects correspond to larger reference variance and lower directional precision.

The causal geometry is incorporated into the flow actor through

$$dX_\tau = u_\theta(s, \tau, X_\tau)d\tau + \sigma V_C^{1/2} dW_\tau, \tag{11}$$

where u_θ is the learned drift field and V_C defines the anisotropic stochastic reference process.

Since the terminal density of a stochastic flow is generally intractable, we follow the path-space formulation of FLAC (Lv et al., 2026) and replace explicit entropy evaluation with a relative-entropy regularization between the policy process and its reference process:

$$\inf_{\mathbb{P}} \left\{ \mathbb{E}_{\mathbb{P}}[\mathcal{G}_s(X_1)] + \lambda D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{P}_C^{\mathrm{ref}}) \right\}, \tag{12}$$

where $\mathbb{P}_C^{\mathrm{ref}}$ is obtained from Eq. (11) by setting $u_\theta = 0$, and the terminal potential is defined as

$$\mathcal{G}_s(x) = -Q_\phi(s, T(x)). \tag{13}$$

Proposition 1 (Optimal endpoint distribution)**.** *For a fixed state s, the optimal endpoint density of Eq.* (12) *is*

$$q_C^\star(x|s) = \frac{1}{Z_C(s)} \mu_{1,C}^{\mathrm{ref}}(x) \exp\left(\frac{Q_\phi(s, T(x))}{\lambda}\right), \tag{14}$$

where $\mu_{1,C}^{\mathrm{ref}}$ is the terminal marginal of the causality-aware reference process. The proof is provided in Appendix A.1.

This formulation separates action preference and exploration geometry: the critic determines high-value regions, while the reference process defines the geometry over which these regions are explored.

Moreover, since the controlled and reference processes share the same diffusion coefficient, Girsanov's theorem gives

$$D_{\mathrm{KL}}(\mathbb{P}^\theta \| \mathbb{P}_C^{\mathrm{ref}}) = \frac{1}{2\sigma^2}\mathbb{E}\int_0^1 u_\theta^\top P_C u_\theta d\tau. \tag{15}$$

Therefore, the corresponding causal kinetic energy is

$$\mathcal{E}_{C,\theta}(s) = \frac{1}{2}\mathbb{E}\int_0^1 u_\theta^\top P_C u_\theta d\tau. \tag{16}$$

The resulting actor objective becomes

$$\mathcal{L}_{\mathrm{actor}}^C = \mathbb{E}\left[-Q_\phi(s, T(X_1)) + \alpha \mathcal{E}_{C,\theta}(s)\right]. \tag{17}$$

Thus, CAFRL introduces a causality-aware path-space regularization: causal effects define the reference geometry and directional kinetic cost, while the critic determines task-oriented action preference.

4.2 CAUSALITY-AWARE ACTOR–CRITIC LEARNING

The previous section establishes that the path-space KL divergence of the causality-aware flow actor is equivalent to the causal kinetic energy $\mathcal{E}_{C,\theta}$. We incorporate this path cost into a SAC-style actor–critic framework with $\lambda = \alpha\sigma^2$.

For a fixed policy and causal geometry, the regularized value is defined as

$$V^\theta(s) = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} \gamma^t \left(R_t - \alpha \mathcal{E}_{C,\theta}(S_t) \right) \middle| S_0 = s \right]. \tag{18}$$

Accordingly,

$$Q^\theta(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} V^\theta(s'), \tag{19}$$

where the critic evaluates task utility and the causal kinetic energy regularizes the path cost of action generation. Since $\mathcal{E}_{C,\theta}$ is independent of the candidate value function, the corresponding Bellman operator remains a γ-contraction; the proof is provided in Appendix A.2.

Actor update. The flow actor generates $A = T(X_N)$ through Eq. (11). Its objective is

$$\mathcal{L}^C_{\text{actor}} = \mathbb{E}_{s,\xi} \left[- \min_{j \in \{1,2\}} Q_{\phi_j}(s, T(X_N)) + \alpha \widehat{\mathcal{E}}^h_{C,\theta}(s) \right], \tag{20}$$

where $\widehat{\mathcal{E}}^h_{C,\theta}$ is the discrete kinetic energy estimator.

Critic update. For replay transitions (S_t, A_t, R_t, S_{t+1}), the target is

$$y^C_t = R_t + \gamma M_t \left[\min_j \bar{Q}_{\bar{\phi}_j}(S_{t+1}, A_{t+1}) - \alpha \widehat{\mathcal{E}}^h_{C,\theta}(S_{t+1}) \right], \tag{21}$$

where A_{t+1} is sampled from the causality-aware flow actor. The critic minimizes

$$\mathcal{L}^C_{\text{critic}} = \mathbb{E} \sum_{j=1}^{2} \left(Q_{\phi_j}(S_t, A_t) - \text{sg}(y^C_t) \right)^2. \tag{22}$$

Together, the critic determines task-oriented action preference, while the causal geometry controls the stochastic exploration scale and kinetic cost of the flow actor.

Theorem 1 (Stability of causality-aware flow sampling)**.** *Assume the causal precision matrices satisfy*

$$0 < p_{\min} I \preceq P^\circ_C, \widehat{P}_C \preceq p_{\max} I,$$

and the drift fields and terminal action map satisfy standard Lipschitz conditions. Let ε_P and ε_u denote the causal geometry and drift approximation errors, respectively. Then the Euler–Maruyama terminal policy distribution satisfies

$$W_2(\pi_h, \pi^\circ) \le C \left(h^{1/2} + \varepsilon_P + \varepsilon_u \right), \tag{23}$$

where C is independent of h, ε_P, and ε_u.

The theorem guarantees that errors in causal geometry estimation, drift approximation, and numerical discretization lead to bounded perturbations of the terminal action distribution. The proof is provided in Appendix A.3.

5 EXPERIMENTS ON HUMANOIDBENCH

We evaluate CAFRL on eight HumanoidBench tasks using the H1Hand robot. Its high-dimensional action space requires coordinated control of the whole body and hands, making efficient exploration particularly challenging. CAFRL addresses this setting by using estimated causal relevance to redistribute action sampling across dimensions, allowing greater exploration along task-relevant directions while constraining less relevant ones.

Our experiments address two questions:

Table 1: Late-training returns on eight HumanoidBench tasks. For each seed, returns are averaged over the final ten evaluations; entries report the mean ± sample standard deviation across five independent training seeds. Bold denotes the highest mean for each task.

Task	FLAC	ACE	SimbaV2	CAFRL
Walk	204.5 ± 46.7	12.4 ± 2.0	118.8 ± 56.2	$\mathbf{699.8} \pm 152.0$
Run	79.1 ± 19.0	7.2 ± 1.5	29.3 ± 4.6	$\mathbf{157.0} \pm 59.3$
Hurdle	44.9 ± 11.1	10.2 ± 0.8	21.0 ± 4.5	$\mathbf{75.2} \pm 6.9$
Stair	70.4 ± 5.8	26.5 ± 1.2	113.1 ± 24.3	$\mathbf{218.3} \pm 80.2$
Balance-hard	78.6 ± 1.5	51.0 ± 1.1	94.0 ± 4.3	$\mathbf{96.9} \pm 8.6$
Crawl	760.8 ± 63.7	578.4 ± 9.3	864.3 ± 90.6	$\mathbf{929.8} \pm 17.0$
Slide	127.1 ± 23.4	22.0 ± 5.0	120.3 ± 16.7	$\mathbf{185.9} \pm 51.9$
Pole	160.0 ± 20.6	54.6 ± 25.3	250.2 ± 77.2	$\mathbf{709.7} \pm 101.6$

- **Q1:** How does CAFRL compare with strong reinforcement-learning baselines in terms of learning efficiency and late-training performance?
- **Q2:** How does the strength of causal anisotropy affect policy learning?

We compare CAFRL against three strong and complementary baselines: FLAC, a flow-based reinforcement-learning method with isotropic path-space regularization (Lv et al., 2026); ACE, a causality-aware maximum-entropy method that incorporates action–reward effects into entropy regularization (Ji et al., 2024); and SimbaV2, a strong model-free reinforcement-learning baseline (Lee et al., 2025). Experiments are conducted on Walk, Run, Hurdle, Stair, Balance-hard, Crawl, Slide, and Pole from HumanoidBench (Sferrazza et al., 2024).

5.1 Main Results

All methods are trained with five independent seeds (0–4) under a common budget of one million environment steps. Checkpoints are evaluated every $10{,}000$ steps using five episodes. For late-training performance, we average the final ten evaluations ($910{,}000$–$1{,}000{,}000$ steps) within each seed and report the mean and sample standard deviation across the five seeds. Detailed configurations are provided in Section C.

Performance across tasks. CAFRL achieves the highest late-training mean return on all eight tasks (Table 1 and Figure 2). The gains are particularly pronounced on Walk, Run, Stair, and Pole. On Walk and Pole, CAFRL reaches 699.8 ± 152.0 and 709.7 ± 101.6, respectively, compared with the strongest baseline means of 204.5 and 250.2. Across all eight tasks, CAFRL achieves a median relative improvement of 80.3% over the strongest baseline. CAFRL also maintains the highest mean return on Hurdle, Slide, Crawl, and Balance-hard, although the performance margins are smaller on the latter two tasks. Overall, the results show that causality-aware anisotropic sampling remains effective across tasks with substantially different locomotion, balance, and whole-body coordination requirements.

Learning efficiency. The learning curves show that the advantage of CAFRL emerges during training rather than only at the final checkpoint. In particular, CAFRL separates clearly from the baselines on Walk, Run, Stair, and Pole as training progresses. This behavior is consistent with the motivation of CAFRL: in the high-dimensional H1Hand action space, causal relevance redistributes exploration toward more relevant action directions instead of assigning the same sampling geometry to every dimension.

5.2 Ablation Studies

Sensitivity to Causal-Anisotropy Strength. We investigate how the strength of causal geometry affects policy learning by varying the exponent η in the relevance-to-precision mapping:

$$p_i = b_C \bar{c}_i^{-\eta}. \tag{24}$$

The parameter η controls the degree of anisotropy by adjusting the contrast of geometry coefficients across action dimensions with different causal relevance.

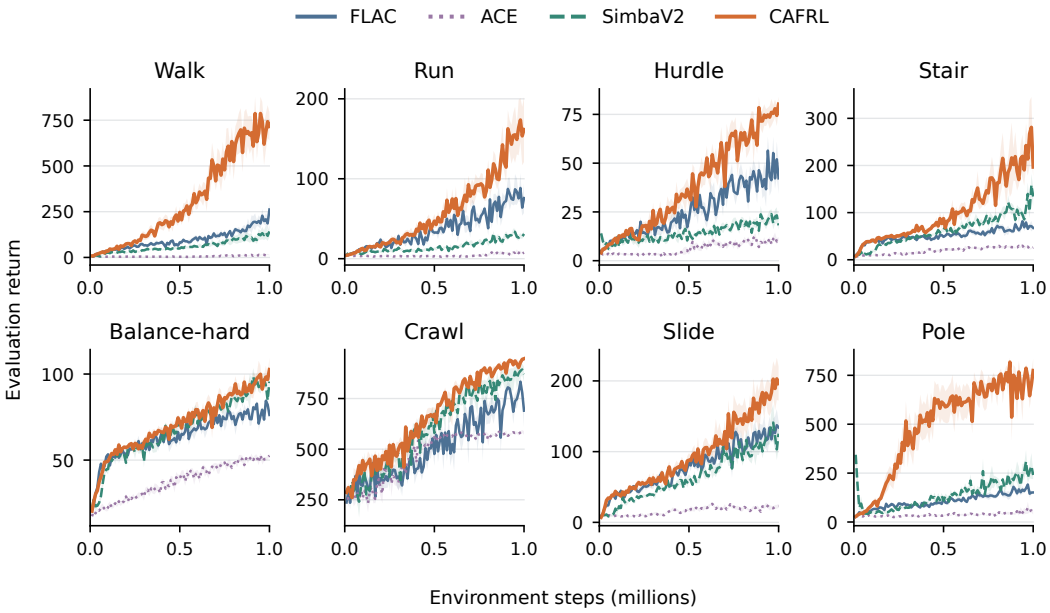


Figure 2: Learning curves over one million environment steps. Lines show the mean across five training seeds and shaded regions show one standard error. Each checkpoint is evaluated over five episodes. Curves are unsmoothed and vertical scales differ across tasks.

Table 2: Walk-task sensitivity analysis over training seeds 0, 1, and 2. Returns are first aggregated within each seed and then reported as mean ± sample standard deviation across seeds. Threshold steps are listed in seed order 0/1/2; > 1.00 indicates that return 700 is not reached within one million environment steps. $0^\dagger$ denotes the isotropic reference without causal weighting. Bold indicates the highest observed mean value.

Exponent η	Average eval. return ↑	Steps to 700 (10^6)	Late return ↑
$0^\dagger$	95.1 ± 13.8	> 1.00 / > 1.00 / > 1.00	198.1 ± 31.0
1.5	191.4 ± 97.3	0.75 / > 1.00 / > 1.00	530.6 ± 308.9
2.5	212.4 ± 95.7	0.80 / > 1.00 / 0.78	580.1 ± 251.8
3.5	290.3 ± 52.7	0.75 / 0.90 / 0.66	$\mathbf{806.9 \pm 46.5}$
5.0	308.2 ± 93.9	0.65 / 0.64 / 0.82	735.9 ± 25.9
6.5	308.0 ± 22.4	0.65 / 0.61 / 0.68	713.6 ± 173.2
7.5	$\mathbf{386.2 \pm 45.8}$	0.49 / 0.50 / 0.52	709.4 ± 179.6

We evaluate $\eta \in \{1.5, 2.5, 3.5, 5.0, 6.5, 7.5\}$ on the Walk task using three training seeds, while keeping the causal precision bounds, mean precision, kinetic-energy scale, and network architecture unchanged. We additionally include an isotropic reference ($0^\dagger$), where all action dimensions share identical geometry coefficients without causal weighting. Training efficiency is measured by the average evaluation return over training checkpoints.

As shown in Table 2 and Figure 3, causal anisotropy consistently improves learning compared with the isotropic reference. Increasing η from 1.5 to 5.0 improves both average training performance and convergence speed, indicating that moderate causal modulation provides a more effective exploration geometry.

The influence of η is not strictly monotonic. Larger values further increase the contrast between causal dimensions, which can improve final performance but may also concentrate exploration more strongly. Among the evaluated settings, $\eta = 7.5$ achieves the highest average return, while intermediate values provide comparable late-stage performance. This suggests that the anisotropy strength controls the trade-off between causal guidance and exploration flexibility.

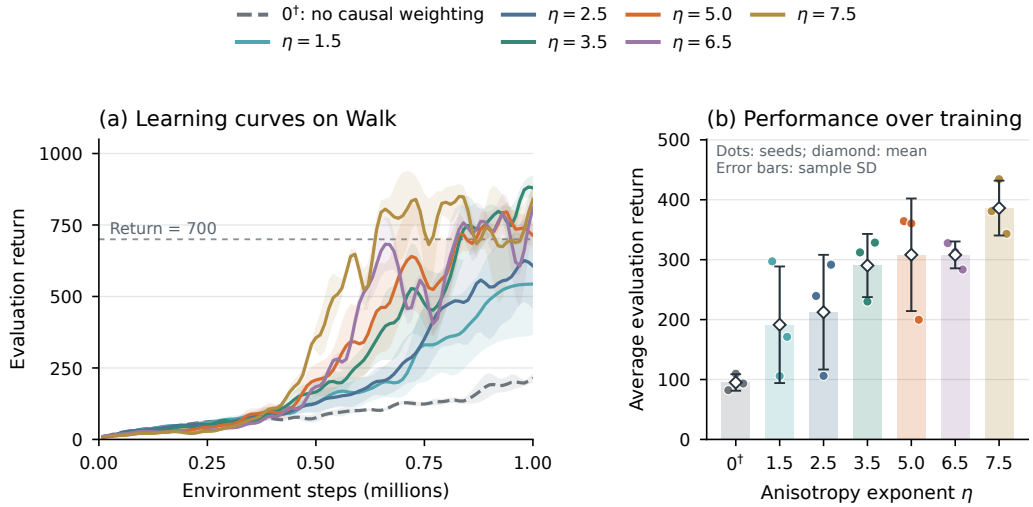


Figure 3: Sensitivity to causal-anisotropy strength on the Walk task. (a) Evaluation learning curves for different values of η; the dashed horizontal line indicates return 700. (b) Average evaluation return over training computed from the original evaluation checkpoints. $0^{\dagger}$ denotes the isotropic reference without causal weighting.

Overall, the ablation validates that causal effects provide a useful signal for shaping the flow reference geometry. Proper anisotropic scaling improves policy learning while maintaining stable flow-based action generation.

6 Conclusion

We introduced CAFRL, a causality-aware flow-based reinforcement-learning framework that incorporates estimated task-relevant causal effects into the reference geometry of policy generation. CAFRL establishes a task-relevant causal policy–reward framework that uses estimated causal effects to characterize action relevance and guide policy generation. By mapping these effects to an anisotropic reference diffusion, CAFRL couples direction-dependent action sampling with causality-aware kinetic regularization through path-space relative entropy. Our analysis characterizes the resulting reference geometry and the stability of the terminal policy under causal-estimation, drift-approximation, and numerical-integration errors. Experiments on eight Humanoid-Bench H1Hand tasks show that CAFRL consistently outperforms strong flow-based, causality-aware, and model-free baselines, while the causal-anisotropy ablation demonstrates that the strength of the induced geometry directly affects policy learning. A remaining limitation is that the current causal estimator relies on linear structural assumptions and observed-state adjustment; extending CAFRL to nonlinear, state-conditional, and latent-confounded causal models is an important direction for future work.

Reproducibility Statement

To facilitate reproducibility, we provide detailed descriptions of the proposed algorithm, optimization objectives, training procedures, and experimental settings in the manuscript and supplementary materials. We report the implementation details, hyperparameter configurations, benchmark environments, and evaluation protocols used in our experiments. All experiments are conducted using publicly available benchmark tasks and standard evaluation procedures. The source code and additional materials will be released upon acceptance to further support reproducibility and future research.

AI Assistance Statement

Large language models (LLMs) were used solely for proofreading and improving the clarity of the manuscript. All scientific contributions, including the proposed methodology, theoretical analysis, experimental design, implementation, and interpretation of results, were conducted and verified by the authors.

References

M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:2160–2169, 2019.

T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:1352–1361, 2017.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80:1861–1870, 2018.

D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.

T. Ji, Y. Liang, Y. Zeng, Y. Luo, G. Xu, J. Guo, R. Zheng, F. Huang, F. Sun, and H. Xu. ACE: Off-policy actor-critic with causality-aware entropy regularization. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:21620–21647, 2024.

H. Lee, Y. Lee, T. Seno, D. Kim, P. Stone, and J. Choo. Hyperspherical normalization for scalable deep reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:33352–33403, 2025.

C. Léonard. Girsanov theory under a finite entropy condition. In *Séminaire de Probabilités XLIV*, Lecture Notes in Mathematics 2046, pp. 429–465, 2012.

C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

G.-H. Liu, Y. Lipman, M. Nickel, B. Karrer, E. Theodorou, and R. T. Q. Chen. Generalized Schrödinger bridge matching. In *International Conference on Learning Representations*, 2024.

L. Lv, Y. Li, Y. Luo, F. Sun, and X. Ma. FLAC: Maximum entropy RL via kinetic energy regularized bridge matching. In *International Conference on Machine Learning*, 2026.

D. McAllister, S. Ge, B. Yi, C. M. Kim, E. Weber, H. Choi, H. Feng, and A. Kanazawa. Flow matching policy gradients. In *International Conference on Learning Representations*, 2026.

M. Seitzer, B. Schölkopf, and G. Martius. Causal influence detection for improving efficiency in reinforcement learning. In *Advances in Neural Information Processing Systems*, 34, 2021.

C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel. HumanoidBench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In *Proceedings of Robotics: Science and Systems XX*, 2024. doi:10.15607/RSS.2024.XX.061.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

Y. Wei, C. Zuo, and Y. Sui. Scalable exploration for high-dimensional continuous control via value-guided flow. In *International Conference on Learning Representations*, 2026.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 34, 2021.

Onno Eberhard, Jakob Hollenstein, Cristina Pinneri, and Georg Martius. Pink noise is all you need: Colored noise exploration in deep reinforcement learning. In *International Conference on Learning Representations*, 2023.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024.

S. Yang, B. Liu, and H. Chen. PolicyFlow: Policy optimization with continuous normalizing flow in reinforcement learning. In *International Conference on Learning Representations*, 2026.

A Theoretical Proofs

A.1 Proof of Proposition 1

Recall the causality-aware reference process

$$dX_\tau^{\mathrm{ref}} = \sigma V_C^{1/2} dW_\tau, \tag{25}$$

and the controlled action-generation process

$$dX_\tau^\theta = u_\theta(s, \tau, X_\tau^\theta)d\tau + \sigma V_C^{1/2} dW_\tau, \tag{26}$$

with the same initial distribution, where

$$V_C = P_C^{-1}. \tag{27}$$

Define

$$\Sigma_C = \sigma V_C^{1/2}. \tag{28}$$

The normalized drift difference between the controlled and reference processes is

$$\vartheta_\tau = \Sigma_C^{-1} u_\theta(s, \tau, X_\tau^\theta) = \frac{1}{\sigma} V_C^{-1/2} u_\theta(s, \tau, X_\tau^\theta). \tag{29}$$

Assume the standard absolute-continuity and integrability conditions required by Girsanov's theorem. In particular, it is sufficient that Novikov's condition holds:

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^1 \|\vartheta_\tau\|_2^2 d\tau\right)\right] < \infty. \tag{30}$$

Girsanov's theorem then gives the path-space relative entropy

$$D_{\mathrm{KL}}\left(\mathbb{P}^\theta \| \mathbb{P}_C^{\mathrm{ref}}\right) = \frac{1}{2}\mathbb{E}_{\mathbb{P}^\theta}\int_0^1 \|\vartheta_\tau\|_2^2 d\tau. \tag{31}$$

Substituting the expression for ϑ_τ yields

$$D_{\mathrm{KL}}\left(\mathbb{P}^\theta \| \mathbb{P}_C^{\mathrm{ref}}\right) = \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}^\theta}\int_0^1 u_\theta^\top V_C^{-1} u_\theta \, d\tau \tag{32}$$

$$= \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}^\theta}\int_0^1 u_\theta^\top P_C u_\theta \, d\tau. \tag{33}$$

By the definition

$$\mathcal{E}_{C,\theta}(s) = \frac{1}{2}\mathbb{E}_{\mathbb{P}^\theta}\int_0^1 u_\theta^\top P_C u_\theta \, d\tau, \tag{34}$$

we obtain

$$\boxed{D_{\mathrm{KL}}\left(\mathbb{P}^\theta \| \mathbb{P}_C^{\mathrm{ref}}\right) = \frac{1}{\sigma^2}\mathcal{E}_{C,\theta}(s)}. \tag{35}$$

This establishes that the causal kinetic-energy regularizer is induced by the path-space relative entropy with respect to the causality-aware reference diffusion.

A.2 Proof of Bellman Contraction

For fixed policy parameters θ, causal geometry P_C, and regularization coefficient α, the causal kinetic energy $\mathcal{E}_{C,\theta}(s)$ does not depend on the argument Q of the Bellman evaluation operator.

For any two bounded action-value functions Q and $\widetilde{Q}$, we therefore have

$$\begin{aligned}
&\left|(\mathcal{T}_C^\theta Q)(s,a) - (\mathcal{T}_C^\theta \widetilde{Q})(s,a)\right| \\
&= \gamma \left|\mathbb{E}_{s'\sim p(\cdot|s,a)}\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[Q(s',a') - \widetilde{Q}(s',a')\right]\right| \\
&\le \gamma \mathbb{E}_{s'\sim p(\cdot|s,a)}\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left|Q(s',a') - \widetilde{Q}(s',a')\right| \\
&\le \gamma \|Q - \widetilde{Q}\|_\infty.
\end{aligned} \tag{36}$$

Taking the supremum over (s,a) gives

$$\boxed{\|\mathcal{T}_C^\theta Q - \mathcal{T}_C^\theta \widetilde{Q}\|_\infty \le \gamma \|Q - \widetilde{Q}\|_\infty}. \tag{37}$$

Hence, for fixed θ, P_C, and α, $\mathcal{T}_C^\theta$ is a γ-contraction. This result concerns exact fixed-policy evaluation and does not imply convergence of the alternating nonlinear actor–critic updates.

A.3 Proof of Causal-Anisotropic Sampling Stability

Let S_C° and $\widehat{S}_C$ denote the corresponding diffusion geometries.

$$S_C^\circ = (P_C^\circ)^{-1/2}, \qquad \widehat{S}_C = (\widehat{P}_C)^{-1/2} \tag{38}$$

denote the ideal and estimated diffusion geometries.

Consider the ideal continuous-time process

$$\begin{aligned}
X_0^\circ &= \sigma_b S_C^\circ \xi_b, \\
dX_\tau^\circ &= u_\tau^\circ(X_\tau^\circ)d\tau + \sigma S_C^\circ dW_\tau,
\end{aligned} \tag{39}$$

and the estimated process

$$\begin{aligned}
\widehat{X}_0 &= \sigma_b \widehat{S}_C \xi_b, \\
d\widehat{X}_\tau &= \widehat{u}_\tau(\widehat{X}_\tau)d\tau + \sigma \widehat{S}_C dW_\tau.
\end{aligned} \tag{40}$$

We couple the two processes using the same Gaussian initial noise ξ_b and the same Brownian motion W_τ.

Assume that the causal precision matrices satisfy

$$0 < p_{\min} I \preceq P_C^\circ, \widehat{P}_C \preceq p_{\max} I, \tag{41}$$

and that the drift fields are globally Lipschitz with constant L and have finite second moments.

Define

$$\Delta_\tau = \widehat{X}_\tau - X_\tau^\circ. \tag{42}$$

The initial perturbation satisfies

$$\mathbb{E}\|\Delta_0\|_2^2 = \sigma_b^2 \mathbb{E}\left\|(\widehat{S}_C - S_C^\circ)\xi_b\right\|_2^2 \tag{43}$$

$$\le C\|\widehat{S}_C - S_C^\circ\|_F^2. \tag{44}$$

From Eqs. (39) and (40),

$$d\Delta_\tau = \left[\widehat{u}_\tau(\widehat{X}_\tau) - u_\tau^\circ(X_\tau^\circ)\right]d\tau + \sigma(\widehat{S}_C - S_C^\circ)dW_\tau. \tag{45}$$

Decompose the drift error as

$$\begin{aligned}&\widehat{u}_\tau(\widehat{X}_\tau) - u^\circ_\tau(X^\circ_\tau)\\ &= \left[\widehat{u}_\tau(\widehat{X}_\tau) - u^\circ_\tau(\widehat{X}_\tau)\right] + \left[u^\circ_\tau(\widehat{X}_\tau) - u^\circ_\tau(X^\circ_\tau)\right].\end{aligned} \tag{46}$$

Define the drift-approximation error

$$\varepsilon_u^2 = \int_0^1 \mathbb{E}\left[\|\widehat{u}_\tau(\widehat{X}_\tau) - u^\circ_\tau(\widehat{X}_\tau)\|_2^2\right] d\tau. \tag{47}$$

Using the Lipschitz property of u°, Itô's formula, and standard quadratic inequalities gives

$$\frac{d}{d\tau}\mathbb{E}\|\Delta_\tau\|_2^2 \le C\mathbb{E}\|\Delta_\tau\|_2^2 + C\mathbb{E}\|\widehat{u}_\tau(\widehat{X}_\tau) - u^\circ_\tau(\widehat{X}_\tau)\|_2^2 + C\|\widehat{S}_C - S^\circ_C\|_F^2. \tag{48}$$

Applying Gronwall's inequality and Eq. (44) yields

$$\mathbb{E}\|\widehat{X}_1 - X^\circ_1\|_2^2 \le C\left(\|\widehat{S}_C - S^\circ_C\|_F^2 + \varepsilon_u^2\right). \tag{49}$$

Under the spectral bounds in Eq. (41), the matrix function $P \mapsto P^{-1/2}$ is Lipschitz on the admissible positive-definite set. Hence,

$$\|\widehat{S}_C - S^\circ_C\|_F \le C_P\|\widehat{P}_C - P^\circ_C\|_F. \tag{50}$$

With

$$\varepsilon_P = \|\widehat{P}_C - P^\circ_C\|_F, \tag{51}$$

Eq. (49) becomes

$$\mathbb{E}\|\widehat{X}_1 - X^\circ_1\|_2^2 \le C\left(\varepsilon_P^2 + \varepsilon_u^2\right). \tag{52}$$

Let X_1^h denote the Euler–Maruyama approximation of the estimated process with step size h. Under the same regularity assumptions, the standard strong-error estimate gives

$$\mathbb{E}\|X_1^h - \widehat{X}_1\|_2^2 \le Ch. \tag{53}$$

Combining Eqs. (52) and (53) yields

$$\mathbb{E}\|X_1^h - X^\circ_1\|_2^2 \le C\left(h + \varepsilon_P^2 + \varepsilon_u^2\right). \tag{54}$$

Finally, assume that the terminal action map T is Lipschitz with constant L_T. Since

$$A_h = T(X_1^h), \qquad A^\circ = T(X^\circ_1), \tag{55}$$

the coupling characterization of the 2-Wasserstein distance gives

$$W_2^2(\pi_h, \pi^\circ) \le \mathbb{E}\|T(X_1^h) - T(X^\circ_1)\|_2^2 \tag{56}$$

$$\le L_T^2\mathbb{E}\|X_1^h - X^\circ_1\|_2^2. \tag{57}$$

Therefore,

$$\boxed{W_2(\pi_h, \pi^\circ) \le C\left(h^{1/2} + \varepsilon_P + \varepsilon_u\right)}. \tag{58}$$

Thus, errors in causal-geometry estimation, drift-field approximation, and numerical discretization propagate continuously to the terminal action distribution.

B Causal-Relevance Estimation Details

B.1 Causal Effect Estimation and DirectLiNGAM

Given a batch of interaction data $\{(S_t, A_t, R_t)\}_{t=1}^B$, we estimate the task relevance of each action dimension from the observed action–reward relationships. To reduce the influence of state-dependent

variations, we condition the causal analysis on the observed state variables and construct the variables used for causal estimation as

$$U = (U_1, \ldots, U_{d_A}, Y), \tag{59}$$

where U_i denotes the state-conditioned variation of action dimension i and Y denotes the corresponding reward variation.

We apply DirectLiNGAM to U to estimate the directed structural relationships among action dimensions and reward (Shimizu et al., 2011). The estimated total causal effect from action dimension i to reward is denoted as

$$\widehat{\tau}_i, \tag{60}$$

which is used as the raw causal relevance score.

The effect magnitudes are transformed into a positive relevance profile through normalization and clipping:

$$\widetilde{c}_i = \mathrm{clip}\left(\frac{|\widehat{\tau}_i| + \epsilon}{d_A^{-1} \sum_j |\widehat{\tau}_j| + \epsilon}, \kappa^{-1}, \kappa\right), \tag{61}$$

followed by

$$c_i = \frac{\widetilde{c}_i}{d_A^{-1} \sum_j \widetilde{c}_j}. \tag{62}$$

For online reinforcement learning, we maintain an exponentially smoothed relevance profile:

$$\bar{c}_i \leftarrow (1 - \beta)\bar{c}_i + \beta c_i. \tag{63}$$

The resulting relevance vector $\bar{\boldsymbol{c}}$ is used to construct the causal reference geometry described in Sec. 4.1.

B.2 Causal Identification Assumptions

The causal interpretation of the estimated relevance profile relies on the standard assumptions of DirectLiNGAM, including an acyclic linear structural model and independent non-Gaussian disturbances (Shimizu et al., 2011). The estimated effects should therefore be interpreted as causal relevance under these identification conditions rather than unconditional causal guarantees.

Algorithm 1 CAFRL with Causal-Effect-Based Reference Geometry

Require: Flow actor, twin critics, replay buffer $\mathcal{B}$, causal update interval T_C

for environment steps $t = 1, 2, \ldots$ **do**
  Construct causal reference geometry (P_C, V_C) from relevance profile $\bar{\boldsymbol{c}}$
  Generate action using the causality-aware flow actor and compute the corresponding kinetic regularization term
  Execute action and store $(S_t, A_t, R_t, S_{t+1}, M_t)$ in $\mathcal{B}$
  Update actor and critics using the causal-energy-regularized SAC objective
  if $t \bmod T_C = 0$ **then**
    Sample interaction data from $\mathcal{B}$
    Estimate causal effects using DirectLiNGAM
    Normalize, clip, and smooth the causal relevance profile
  end if
end for

C Experimental Configurations

All methods use the original task rewards, $5{,}000$ initial random-exploration steps, discount factor 0.99, one gradient update per interaction after initialization, and training seeds 0–4. Evaluation uses five episodes every $10{,}000$ steps. We restrict every comparison to the first 10^6 steps, irrespective of a run's configured maximum horizon. No checkpoint is selected by its best evaluation return.

Table 3: Principal method-specific hyperparameters. Widths refer to actor/critic hidden dimensions; the target-update rate is the interpolation coefficient for target-network updates.

Parameter	FLAC	ACE	SimbaV2	CAFRL
Batch size	256	512	256	256
Actor/critic width	512/512	1024/1024	128/512	512/512
Learning rate	3×10^{-4}	3×10^{-4}	$10^{-4}\rightarrow5\times10^{-5}$	3×10^{-4}
Target-update rate	0.1	0.005	0.005	0.1
Kinetic-energy target	2.5	–	–	Task-dependent

CAFRL's kinetic-energy target coefficient is 0.5 on Walk, Run, Hurdle, and Slide; 0.25 on Balance-hard; 0.2 on Stair; and 0.1 on Crawl and Pole. FLAC uses 2.5 on every task. Both use distributional critics with 101 atoms over $[-10, 150]$, whereas SimbaV2 uses its normalized architecture, reward normalization, and 101 atoms over $[-5, 5]$. These settings constitute method comparisons, rather than comparisons with matched architectures, energy targets, or hyperparameter-search budgets.

CAFRL and ACE estimate action relevance from $4{,}096$ replay samples using state-residualized DirectLiNGAM, with causal updates every $50{,}000$ gradient updates. In the main experiment, CAFRL clips normalized relevance weights to $[0.2, 5]$ and maps them to inverse-covariance coefficients bounded by $[0.5, 2]$, using exponent 1.25. ACE retains its own entropy-weighting rule. Its entropy coefficient is automatically tuned on Walk, Run, Hurdle, and Balance-hard, and fixed at 0.02 on Stair, Crawl, Slide, and Pole. Balance-hard uses an entropy target of 22 and initial coefficient 0.02; ACE network resets occur every $100{,}000$ steps. The released run manifest records all configurations and exact run identifiers, including the selected task-specific CAFRL settings and the three designated Balance-hard baseline groups.

The exponent ablation in Section 5.2 is conducted on the Walk task using three training seeds (0,1,2), with precision bounds $[0.00625, 80]$, energy target 0.5, and $\eta \in \{1.5, 2.5, 3.5, 5.0, 6.5, 7.5\}$. The causal-weight update coefficient is set to 1, such that each update directly replaces the previous relevance profile. All other training and evaluation settings follow the main configuration. The isotropic reference ($0^{\dagger}$) uses the same evaluation protocol but removes causal weighting; it is included to quantify the effect of causal anisotropy.