CADENT: Gated Hybrid Distillation for Sample-Efficient Transfer in Reinforcement Learning

Mahyar Alinejad[1] **Yue Wang**[1,2] **George K. Atia**[1,2]

[1]Dept. of Electrical & Computer Engineering, University of Central Florida [2]Dept. of Computer Science, University of Central Florida

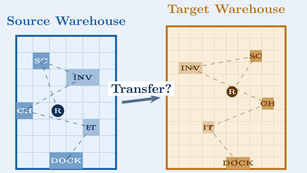
{mahyar.alinejad, yue.wang, george.atia}@ucf.edu

Problem Setup

Deep RL requires **millions of interactions** to learn from scratch. **Transfer learning** reuses a pre-trained **teacher** to accelerate a **student** in a new, related task.

Source MDP: $\mathcal{M}_{src} = (\mathcal{S}_{src}, \mathcal{A}, P_{src}, R_{src}, \gamma)$
Target MDP: $\mathcal{M}_{tgt} = (\mathcal{S}_{tgt}, \mathcal{A}, P_{tgt}, R_{tgt}, \gamma)$

Tasks share $\mathcal{A}$ and γ but differ in states, dynamics, and rewards (**domain shift**). The teacher provides $Q_{teacher}(s,a)$ and a task automaton $\mathcal{D}$; the student must reach π^*_{tgt} as fast as possible.

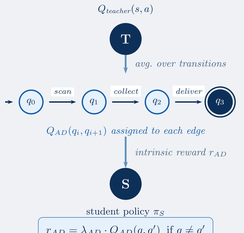


6×8 — *teacher trained here* 10×12 — *student learns here*

SC scanner INV inventory CH charging IT item DOCK shipping dock R robot *dashed* = workflow

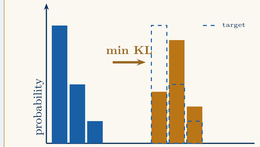
Two Types of Teacher Knowledge

(S) Strategic
Automaton Distillation



Teacher's task structure bootstrapped into student's reward.

(T) Tactical
Policy Distillation



$\mathcal{L}_{PD} = D_{KL}(\pi_T \| \pi_S)$

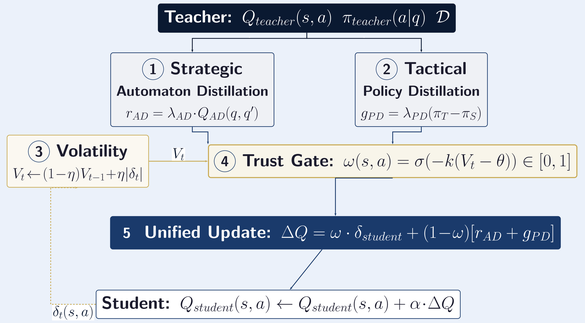
Gradient correction in Q-update:
$g_{PD} = \lambda_{PD}(\pi_T(a|q) - \pi_S(a|s))$

Fine-grained local action guidance at every state.

Gap in Prior Work

Method	Strategic	Tactical	Adaptive
Policy Distillation [1]	×	✓	×
Automaton Distillation [3]	✓	×	×
Kickstarting DRL [2]	×	✓	Global only
CADENT (Ours)	✓	✓	Per (s,a)

CADENT Framework Overview



Early: $V_t > \theta \Rightarrow \omega \to 0$: teacher dominates **Late:** $V_t < \theta \Rightarrow \omega \to 1$: student autonomous

Key Equations

(1) Strategic Guidance via Automaton Distillation

Distil teacher Q-values into per-transition values:

$$Q_{AD}(q,q') = \frac{1}{|\mathcal{T}_{q \to q'}|} \sum_{(s,a) \in \mathcal{T}_{q \to q'}} Q_{teacher}(s,a), \qquad r_{AD} = \begin{cases} \lambda_{AD} \cdot Q_{AD}(q,q') & q \neq q' \\ 0 & \text{otherwise} \end{cases}$$

(2) Tactical Guidance via Policy Distillation

Bounded policy-gradient correction nudging student toward teacher:

$$g_{PD}(s,a) = \lambda_{PD} \cdot \big(\pi_{teacher}(a|q) - \pi_{student}(a|s)\big)$$

(3) & (4) Experience-Gated Trust Mechanism

Per-(s,a) volatility tracker (EMA of TD-errors) and sigmoid gate:

$$V_t(s,a) \leftarrow (1-\eta)\, V_{t-1}(s,a) + \eta\, |\delta_t(s,a)|, \qquad \omega(s,a) = \sigma\big(-k\,(V_t(s,a) - \theta)\big) \in [0,1]$$

(5) Unified Q-Learning Update

$$Q_{student}(s,a) \leftarrow Q_{student}(s,a) + \alpha \Big[\underbrace{\omega(s,a) \cdot \delta_{student}}_{\text{Student experience}} + \underbrace{(1 - \omega(s,a)) \cdot \big[r_{AD} + g_{PD}(s,a)\big]}_{\text{Teacher guidance}} \Big]$$

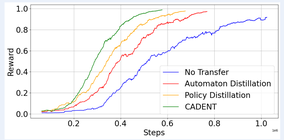
Evaluation Environments

CADENT is evaluated on **four diverse environments** that span fundamentally different domain types, ensuring broad generality of the findings:

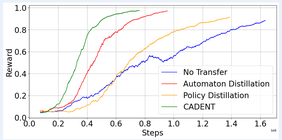
Environment	Challenge
Blind Craftsman	Sparse 25×25 gridworld with looping subgoal structure; severe long-horizon exploration
Dungeon Quest	20×20 maze with strict logical prerequisites; sequential item acquisition
Mountain Car Coll.	9-D physics; agent must learn momentum-building oscillation strategy under resource constraints
Warehouse Robotics	12-D robotics with **explicit domain shift**: teacher on 6×8, student on 10×12 layout with different station arrangements

Sample Efficiency: Reward per Cumulative Steps

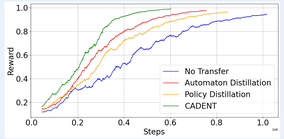
The primary metric directly measures how much environment interaction each method needs to reach a given performance level. **CADENT reaches performance levels that baselines require 40–60% more samples to achieve.**



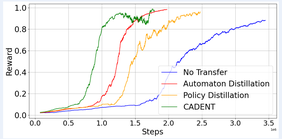
Blind Craftsman



Dungeon Quest



Mountain Car Collection



Warehouse Robotics

Theoretical Guarantee

Proposition (Bounded Updates). With $|R(s,a)| \leq R_{max}$ and $|Q_{AD}(q,q')| \leq Q^{AD}_{max}$:

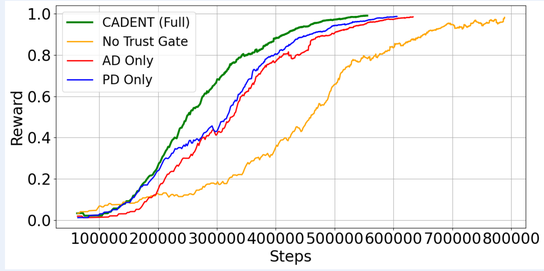
$$|\Delta Q(s,a)| \leq \frac{R_{max}}{1-\gamma} + \lambda_{AD} Q^{AD}_{max} + 2\lambda_{PD}$$

Since $\omega(s,a) \in [0,1]$, the update is a convex combination of bounded terms, **preventing divergence** regardless of trust mechanism state.

Trust properties: (i) *monotone*—trust decreases as volatility increases; (ii) *threshold* at θ; (iii) *graceful degradation* even under poor teacher.

Ablation Study — Blind Craftsman

To isolate each component's contribution, three variants of CADENT are compared by removing one component at a time: **No Trust Gate** ($\omega = 0.5$ fixed, removing adaptation), **AD Only** ($\lambda_{PD} = 0$, no tactical guidance), and **PD Only** ($\lambda_{AD} = 0$, no strategic guidance).



The sample efficiency curve confirms that all three components are **individually necessary**: removing the trust gate causes early over-reliance on the teacher; removing either guidance stream reduces the quality of the learning signal. Only the full CADENT achieves optimal sample efficiency.

Summary & Contributions

CADENT is the first framework to:

1. **Unify** automaton-based *strategic* and policy-based *tactical* guidance into one coherent learning signal.
2. **Gate** teacher influence *per state-action pair* via TD-volatility, outperforming Kickstarting by **+24%** sample efficiency.
3. Achieve **40–60%** better sample efficiency across gridworlds, physics control, and high-dimensional robotics with superior asymptotic performance.

Future: deep function approximation (DQN/PPO), lifelong learning with multiple teachers, formal convergence analysis for adaptive trust systems.

References

- A. Rusu et al. *Policy Distillation.* arXiv:1511.06295, 2015.
- S. Schmitt et al. *Kickstarting Deep RL.* arXiv:1803.03835, 2018.
- S. Singireddy et al. *Automaton Distillation.* arXiv:2310.19137, 2023.
- M. Alinejad et al. *Bidirectional Framework.* PMLR 288:643, 2025.
- R. Icarte et al. *Reward Machines.* JAIR 73:1–44, 2022.

Supported by DARPA HR0011-24-9-0427 & NSF CCF-2106339