Learnable Behavior Control: Breaking Atari Human World Records via Sample-Efficient Behavior Selection

Jiajun Fan, Yuzheng Zhuang, Yuecheng Liu, Jianye HAO,
Bin Wang, Jiangcheng Zhu, Hao Wang, Shu-Tao Xia

Presenter: Jiajun Fan
Email: fanjj21@mails.tsinghua.edu.cn

Introduction

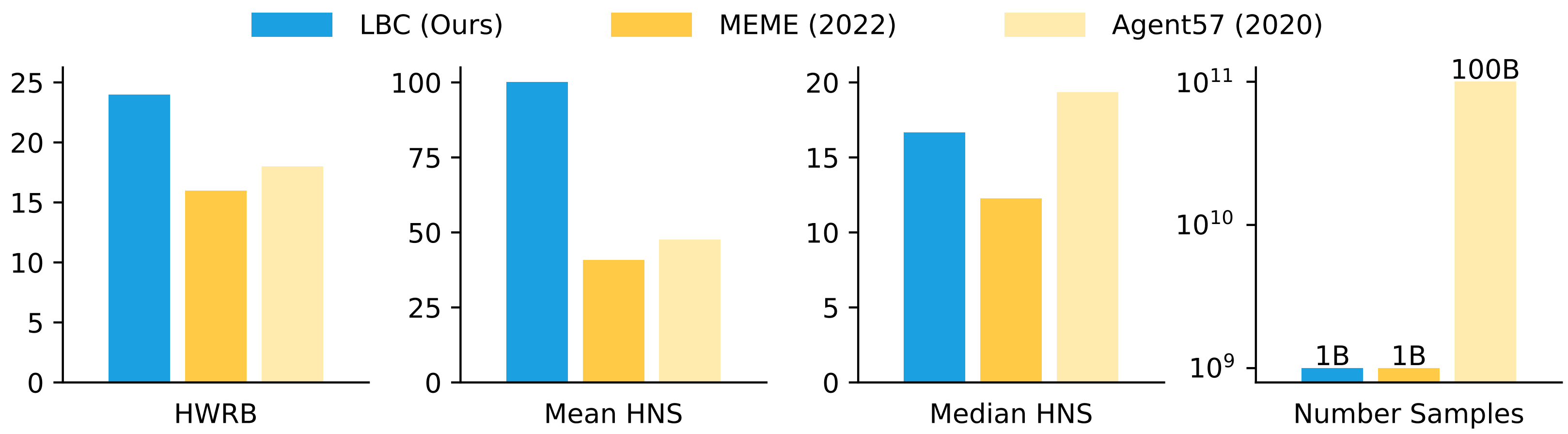

Figure 1: Performance on the Atari.

1. The efficacy of reinforcement learning (RL) algorithms in practical applications is heavily reliant on their sampling efficiency.

2. Achieving optimal performance with limited data samples is a challenging task, and only a handful of algorithms can achieve both high sample efficiency and superior final performance.

3. While some RL models have demonstrated remarkable results in specific tasks, the claim of surpassing human-level performance is often exaggerated and misleading. Despite recent advancements in RL, the strongest algorithms still fall short of outperforming human world records on a multitude of tasks.

Introduction

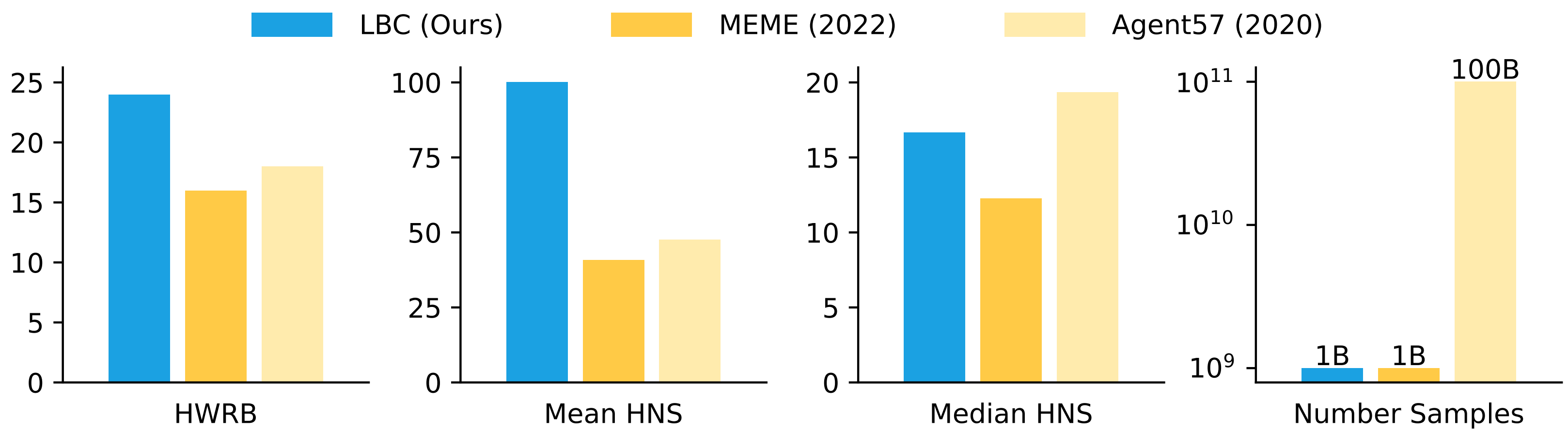

Figure 1: Performance on the Atari.

1. Reduce the amount of **training data** by current SOTA reinforcement learning algorithms by more than **20-100 times**.

2. In the case of reduced data sample size, **maintain or even surpass** SOTA performance, and even surpasses the original performance.

3. Break **all human world records** and obtain **real** super-human agents in Atari.

Why do we need behavior control?

Better data facilitate **better performance** and better sample efficiency.

How to optimize the data distribution in RL?

Behavior Policy!

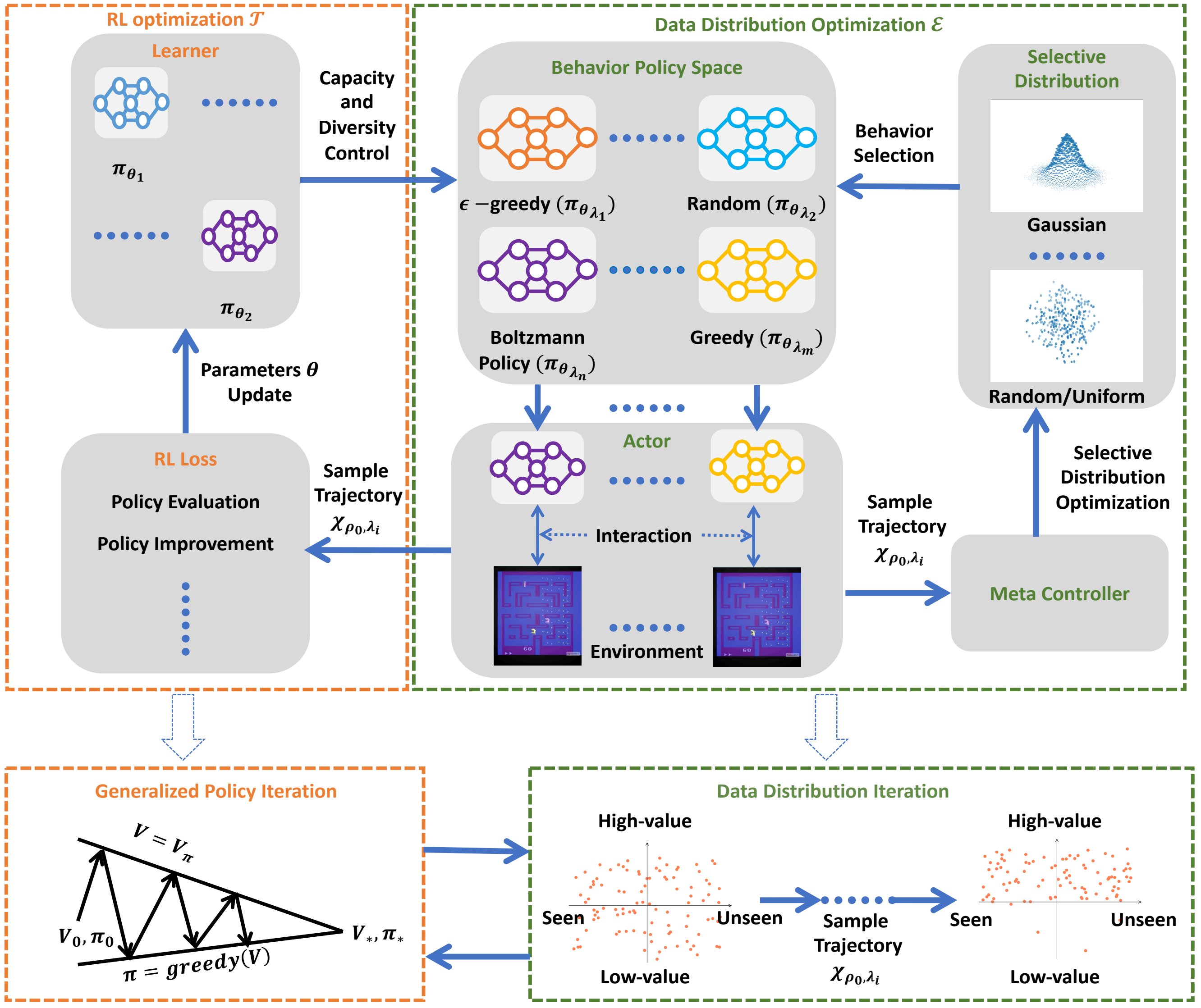

Figure 2: Data Distribution Optimization [1]

[1] Fan, Jiajun, and Changnan Xiao. "Generalized Data Distribution Iteration." *International Conference on Machine Learning*. PMLR, 2022.

Why do we need behavior control?

Theorem 1 (First-Order Optimization with Superior Target). *Under assumptions* (1) (2) (3), *we have* $\mathcal{L}_{\mathcal{T}}(\mathcal{P}_{\Lambda}^{(t+1)}, \theta^{(t+1)}) = \boldsymbol{E}_{\lambda\sim\mathcal{P}_{\Lambda}^{(t+1)}}[\mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1)})] \geq \boldsymbol{E}_{\lambda\sim\mathcal{P}_{\Lambda}^{(t)}}[\mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1)})] = \mathcal{L}_{\mathcal{T}}(\mathcal{P}_{\Lambda}^{(t)}, \theta^{(t+1)})$.

Theorem 2 (Second-Order Optimization with Superior Improvement). *Under assumptions* (1) (2) (4), *we have* $\boldsymbol{E}_{\lambda\sim\mathcal{P}_{\Lambda}^{(t+1)}}[G^{\eta}\mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1)})] \geq \boldsymbol{E}_{\lambda\sim\mathcal{P}_{\Lambda}^{(t)}}[G^{\eta}\mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1)})]$, *more specifically,*

$$\boldsymbol{E}_{\lambda\sim\mathcal{P}_{\Lambda}^{(t+1)}}[\mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1),\eta}) - \mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1)})]$$
$$\geq \boldsymbol{E}_{\lambda\sim\mathcal{P}_{\Lambda}^{(t)}}[\mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1),\eta}) - \mathcal{L}_{\mathcal{T}}(\lambda, \theta_{\lambda}^{(t+1)})]$$

If we use better data for training, can we obtain better performance? Yes!

Better Data facilitate Better RL training!
But How? -> LBC

Behavior Control Formulation

Definition 3.1 (Behavior Space Construction). *Considering the RL problem that behaviors μ are generated from some policy model(s). We can acquire a family of realizable behaviors by applying a family of behavior mappings $\mathcal{F}_{\mathbf{\Psi}}$ to these policy model(s). Define the set that contains all of these realizable behaviors as the behavior space, which can be formulated as:*

$$\mathbf{M}_{\mathbf{\Theta},\mathbf{H},\mathbf{\Psi}} = \begin{cases} \{\mu_{\boldsymbol{\theta},\mathbf{h},\psi} = \mathcal{F}_{\psi}(\Phi_{\mathbf{h}}) | \boldsymbol{\theta} \in \mathbf{\Theta}, \mathbf{h} \in \mathbf{H}, \psi \in \mathbf{\Psi}\}, & \textit{for individual behavior mapping} \\ \{\mu_{\mathbf{\Theta},\mathbf{H},\psi} = \mathcal{F}_{\psi}(\mathbf{\Phi}_{\mathbf{\Theta},\mathbf{H}}) | \psi \in \mathbf{\Psi}\}, & \textit{for hybrid behavior mapping} \end{cases} \tag{2}$$

Assumption 1

$$\mathbf{M}_{\mathbf{H},\mathbf{\Psi}} = \begin{cases} \{\mu_{\mathbf{h},\psi} = \mathcal{F}_{\psi}(\Phi_{\mathbf{h}}) | \mathbf{h} \in \mathbf{H}, \psi \in \mathbf{\Psi}\}, & \text{for individual behavior mapping} \\ \{\mu_{\mathbf{H},\psi} = \mathcal{F}_{\psi}(\mathbf{\Phi}_{\mathbf{H}}) | \psi \in \mathbf{\Psi}\}, & \text{for hybrid behavior mapping} \end{cases} \tag{4}$$

Definition 3.2 (Behavior Selection). *Behavior selection can be formulated as finding a optimal selection distribution $\mathcal{P}^*_{\mathbf{M}_{\mathbf{\Theta},\mathbf{H},\mathbf{\Psi}}}$ to select the behaviors μ from behavior space $\mathbf{M}_{\mathbf{\Theta},\mathbf{H},\mathbf{\Psi}}$ and maximizing some optimization target $\mathcal{L}_{\mathcal{P}}$, wherein $\mathcal{L}_{\mathcal{P}}$ is the optimization target of behavior selection:*

$$\mathcal{P}^*_{\mathbf{M}_{\mathbf{\Theta},\mathbf{H},\mathbf{\Psi}}} := \underset{\mathcal{P}_{\mathbf{M}_{\mathbf{\Theta},\mathbf{H},\mathbf{\Psi}}}}{\operatorname{argmax}} \mathcal{L}_{\mathcal{P}} \tag{3}$$

Assumption 1

$$\mathcal{P}^*_{\mathbf{M}_{\mathbf{H},\mathbf{\Psi}}} := \underset{\mathcal{P}_{\mathbf{M}_{\mathbf{H},\mathbf{\Psi}}}}{\operatorname{argmax}} \mathcal{L}_{\mathcal{P}}$$

Behavior Control Method

Behavior Control

$$\mathbf{M}_{\mathbf{H},\boldsymbol{\Psi}} = \begin{cases} \{\mu_{\mathbf{h},\psi} = \mathcal{F}_{\psi}(\Phi_{\mathbf{h}}) | \mathbf{h} \in \mathbf{H}, \psi \in \boldsymbol{\Psi}\}, & \text{for individual behavior mapping} \\ \{\mu_{\mathbf{H},\psi} = \mathcal{F}_{\psi}(\boldsymbol{\Phi}_{\mathbf{H}}) | \psi \in \boldsymbol{\Psi}\}, & \text{for hybrid behavior mapping} \end{cases}$$

Prop. 1

$$\mathbf{M}_{\mathbf{H},\boldsymbol{\Psi}} = \begin{cases} \{\mu_{\mathbf{h},\psi} = \mathcal{F}_{\psi}(\Phi_{\mathbf{h}}) | \mathbf{h} \in \mathbf{H}, \psi \in \boldsymbol{\Psi}\}, & \text{for individual behavior mapping} \\ \{\mu_{\mathbf{H},\psi} = \mathcal{F}_{\psi}(\boldsymbol{\Phi}_{\mathbf{H}}) | \psi \in \boldsymbol{\Psi}\}, & \text{for hybrid behavior mapping} \end{cases}$$

Prop. 2

Proposition 1 (Policy Model Selection). *When $\mathcal{F}_{\psi}$ is a deterministic and individual behavior mapping for each actor at each training step (wall-clock), e.g., **Agent57**, the behavior for each actor can be uniquely indexed by* $\mathbf{h}$, *so equation 5 can be simplified into*

$$\mathcal{L}_{\mathcal{P}} = \mathbb{E}_{\mathbf{h} \sim \mathcal{P}_{\mathbf{H}}} \left[V_{\mu_{\mathbf{h}}}^{\mathrm{TV}} + c \cdot V_{\mu_{\mathbf{h}}}^{\mathrm{TD}} \right], \tag{6}$$

where $\mathcal{P}_{\mathbf{H}}$ is a selection distribution of $\mathbf{h} \in \mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_{\mathrm{N}}\}$. For each actor, the behavior is generated from a selected policy model $\Phi_{\mathbf{h}_i}$ with a pre-defined behavior mapping $\mathcal{F}_{\psi}$.

Agent57, NGU

Proposition 2 (Behavior Mapping Optimization). *When all the policy models are used to generate each behavior, e.g., $\mu_{\psi} = \mathcal{F}_{\psi}(\Phi_{\boldsymbol{\theta},\mathbf{h}})$ for single policy model cases or $\mu_{\psi} = \mathcal{F}_{\psi}(\Phi_{\boldsymbol{\theta}_1,\mathbf{h}_1}, ..., \Phi_{\boldsymbol{\theta}_{\mathrm{N}},\mathbf{h}_{\mathrm{N}}})$ for* N *policy models cases, each behavior can be uniquely indexed by $\mathcal{F}_{\psi}$, and equation 5 can be simplified into:*

$$\mathcal{L}_{\mathcal{P}} = \mathbb{E}_{\psi \sim \mathcal{P}_{\boldsymbol{\Psi}}} \left[V_{\mu_{\psi}}^{\mathrm{TV}} + c \cdot V_{\mu_{\psi}}^{\mathrm{TD}} \right], \tag{7}$$

where $\mathcal{P}_{\boldsymbol{\Psi}}$ is a selection distribution of $\psi \in \boldsymbol{\Psi}$.

LBC (Ours)

Hybrid Behavior Mapping

1. **Generalized Policy Selection.** Adjusting the contribution proportion of each learned policy for the behavior via an importance weight w.

2. **Policy-Wise Entropy Control.** Controlling the entropy of each policy via an entropy control function f.

3. **Behavior Distillation from Multiple Policies.** Distilling the entropy-controlled policies into a behavior policy according to the proportion of contribution and a behavior distillation function g.

$$\mathbf{M}_{\mathbf{H},\Psi} = \left\{ g\left(f_{\tau_1}(\Phi_{\mathbf{h}_1}), \ldots, f_{\tau_N}(\Phi_{\mathbf{h}_N}), \omega_1, \ldots, \omega_N \right) \middle| \psi \in \Psi \right\}$$

To control the behavior, the only thing we have to do is to optimize $\psi = (\tau_1, \omega_1 \ldots \tau_N, \omega_N) \in \Psi$ with a meta-controller since f, g, N, H are predefined.

Framework

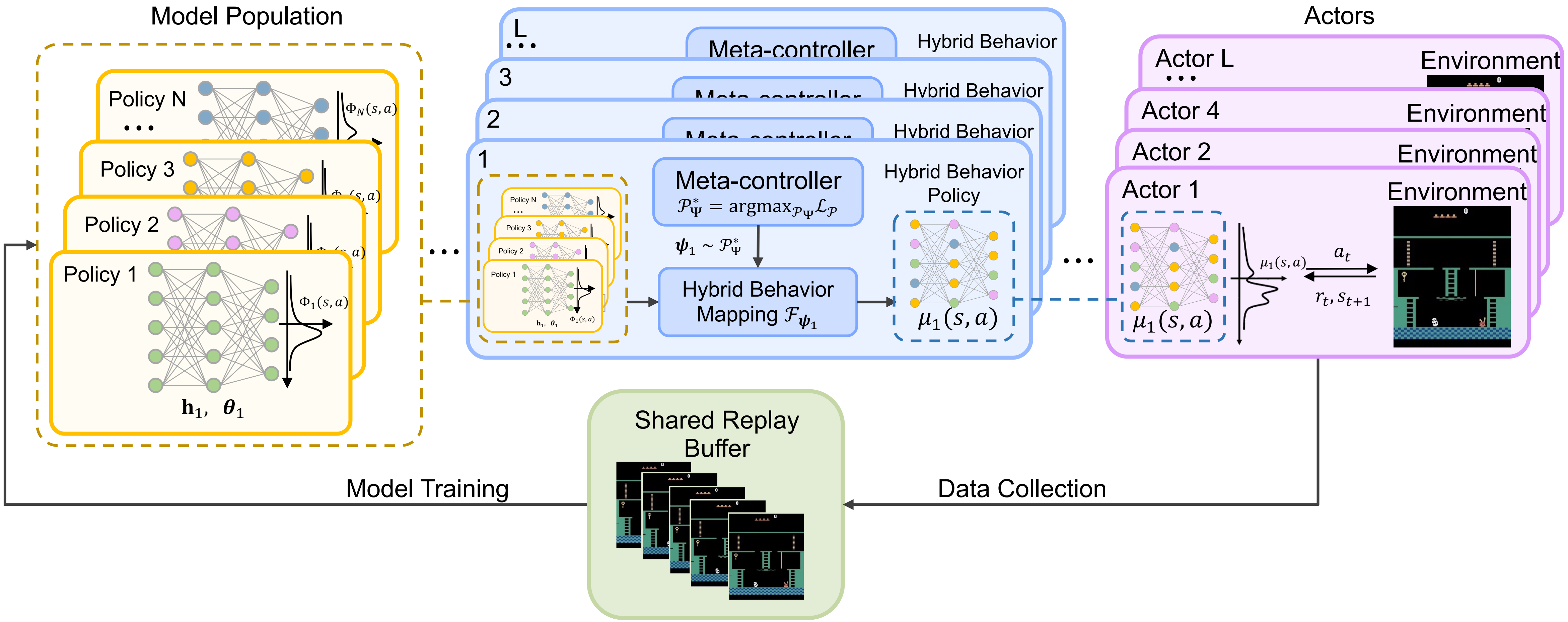

Figure 3: A general framework of LBC.

Experiment

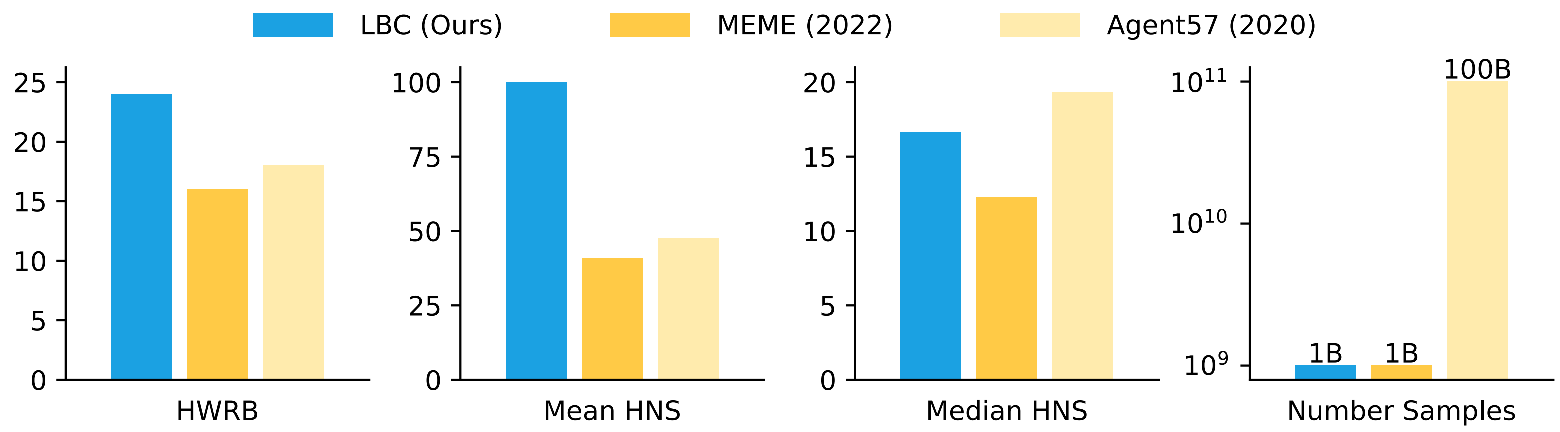

Figure 4: Performance on the 57 Atari.

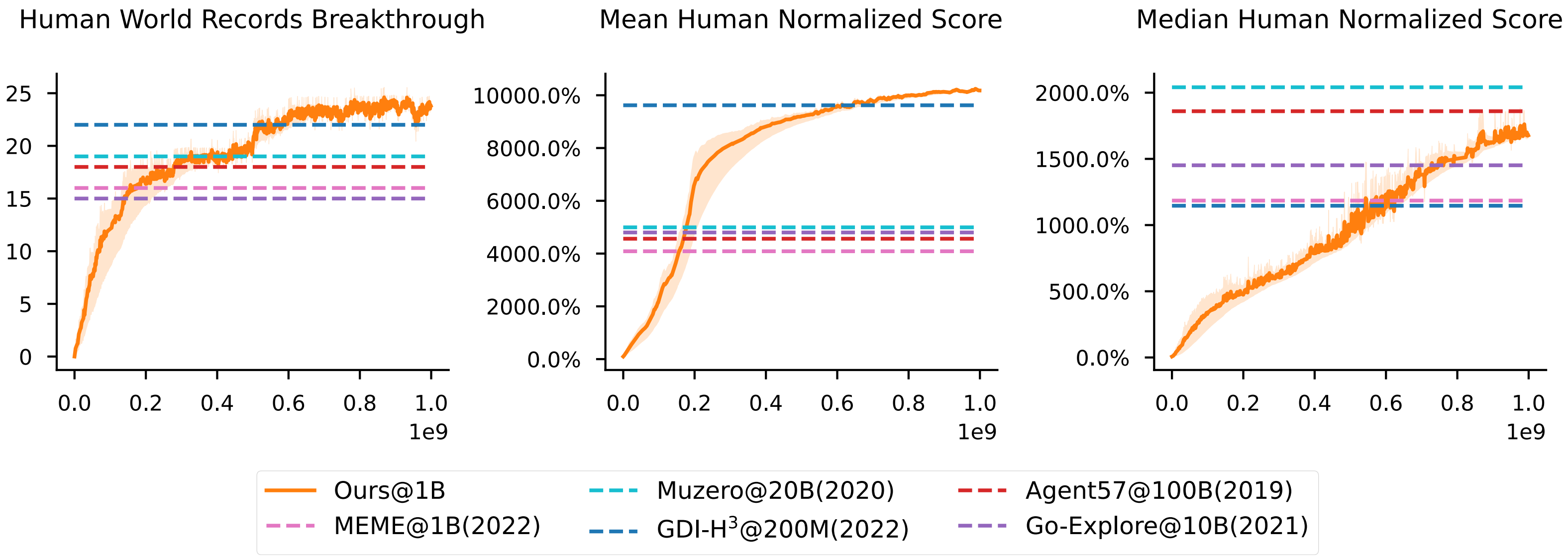

Figure 5: Atari Learning Curve

Experiment

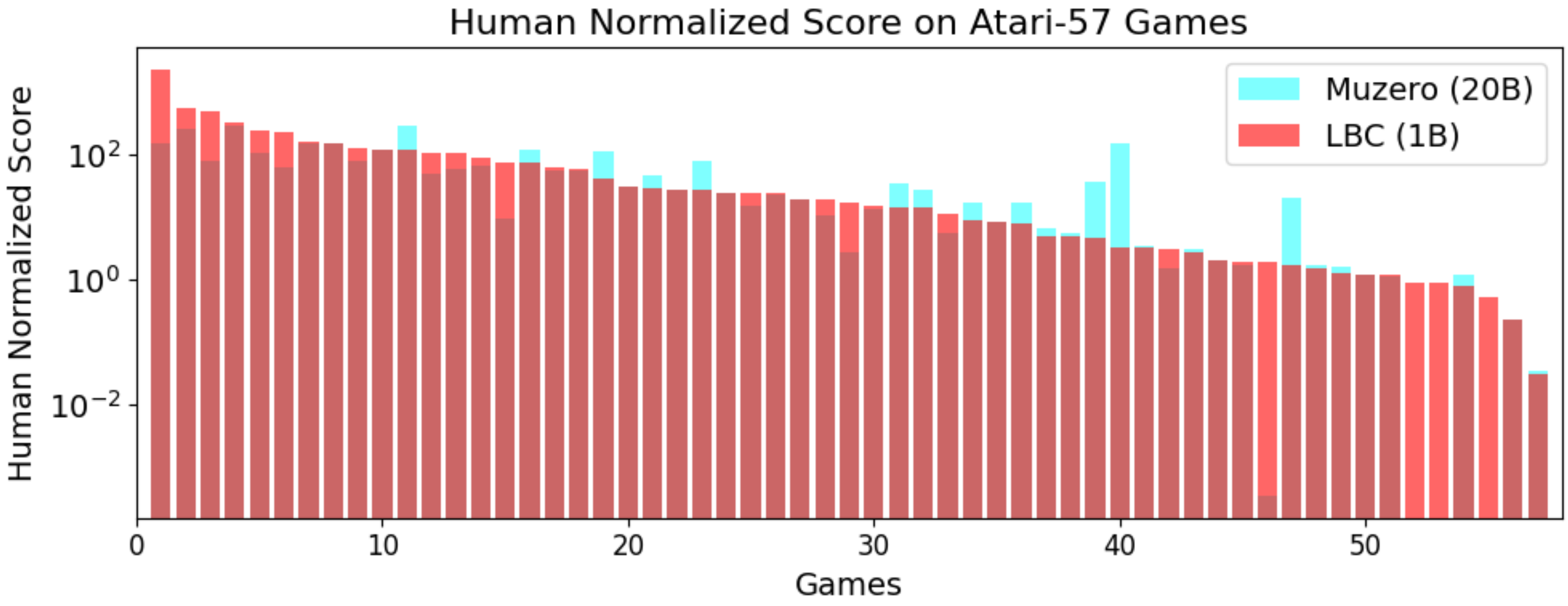

Figure 6: Comparison with Muzero. Human-normalized scores per game at different interaction budgets, sorted from highest to lowest.

Conclusion and Research Map

Behavioral Control in RL

1. GDI: **Theoretical Guarantee**. Behavioral control in single policy RL. (Done)
2. LBC: **General way**. Behavior control in population-based RL. (Done)
3. Multi-Game LBC: Behavior control in **Multi-Task RL**. (In progress)
4. Robo BC: Behavior control in **Robotics**. (In progress)

What’s Next?
Can We Unify the Behavior Control in RL? Yes!

Thank you for your listening!

Contact: fanjj21@mails.tsinghua.edu.cn
Blogs: lbc.jiajunfan.com