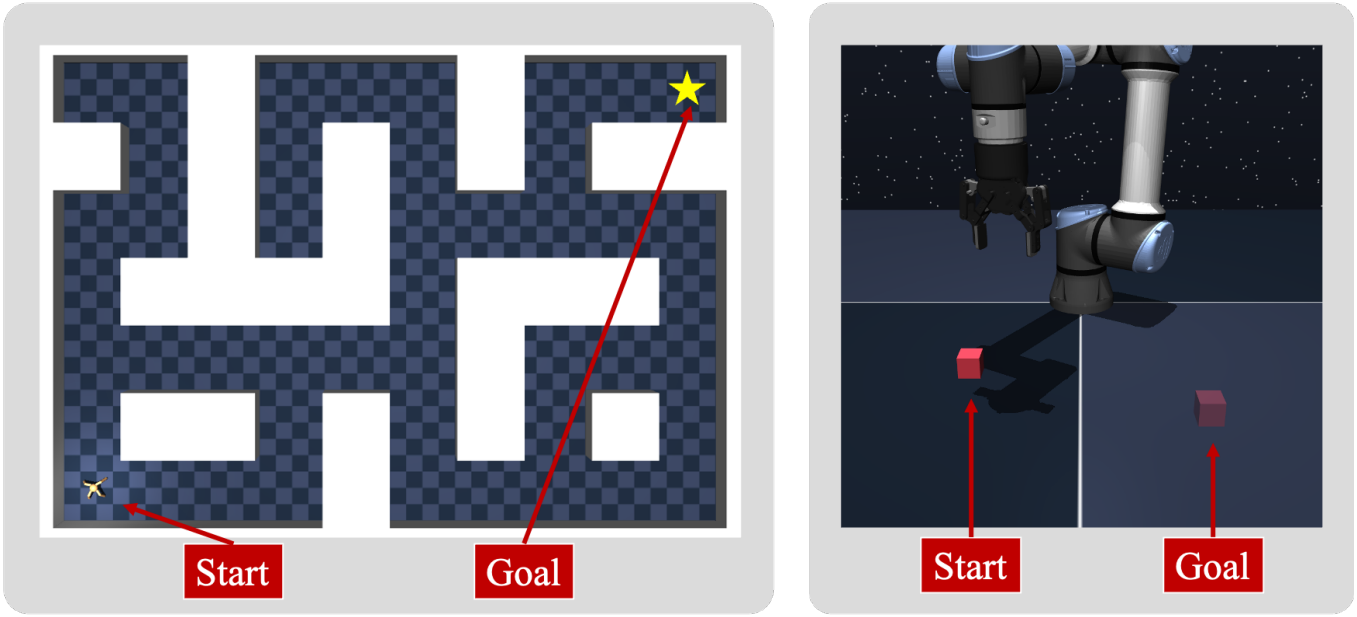
Latent Representation Alignment for Offline Goal-Conditioned Reinforcement Learning

Hyungkyu Kang[*], **Byeongchan Kim**[*], and **Min-hwan Oh**

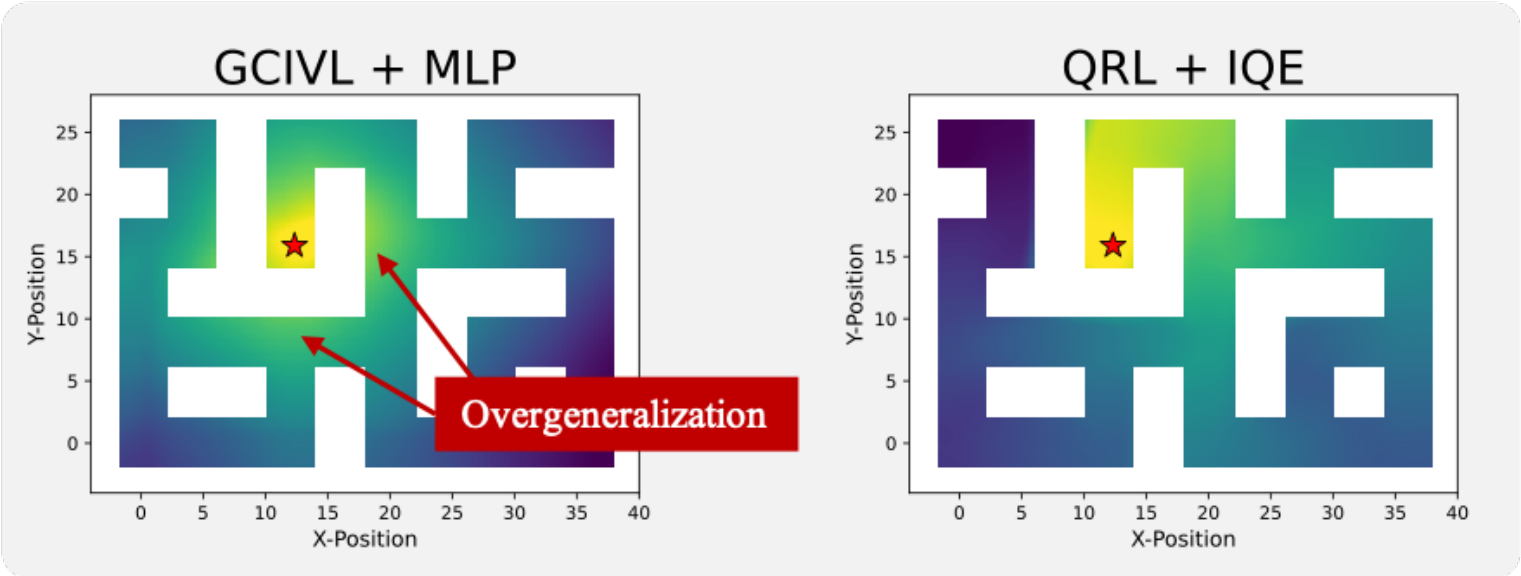
Seoul National University, [*]Equal contribution.

Offline Goal-Conditioned RL (GCRL)



- Discounted Markov decision process (MDP) augmented with goal space $\mathcal{G}$: $(\mathcal{S}, \mathcal{A}, \mathcal{G}, p_0, p, r, \gamma)$
- Reward $r(s, a, g)$ depends on goal $g \in \mathcal{G}$.
- We have access to an offline dataset $\mathcal{D}$ consisting of trajectories $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$ collected by some behavior policy μ.
- Goal: Maximize the expected cumulative reward $J(\pi) := \mathbb{E}_{g \sim q, \tau \sim p^\pi} [\sum_{t=0}^{T} \gamma^t r(s_t, g)]$

Motivation: Bottleneck of GCRL

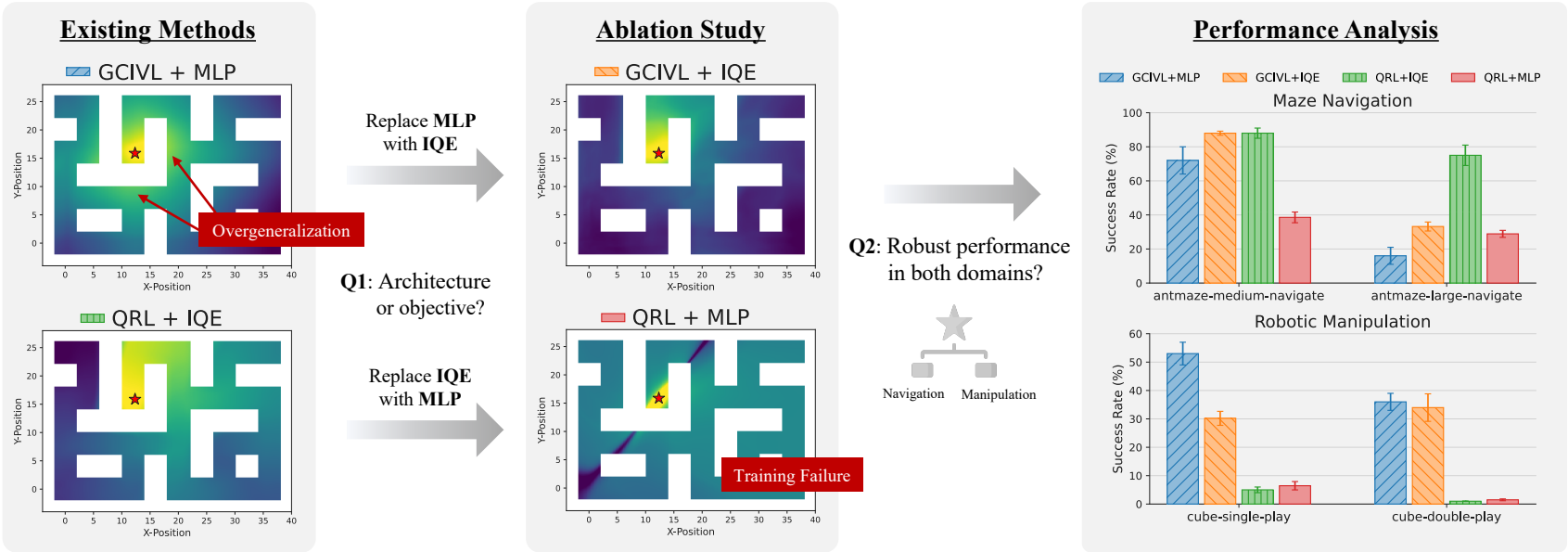


- GCIVL (Kostrikov et al., 2022) is a TD learning algorithm that estimates the optimal value by expectile regression, which uses the MLP value function by default.
- QRL (Wang et al., 2023) computes optimal value by constrained optimization of a quasimetric network such as IQE (Wang and Isola, 2022).

GCIVL+MLP exhibits Overgeneralization, an **erroneous generalization** of value estimate, assigning similar values to states that are close in Euclidean distance, not in temporal distance.

Motivation: Bottleneck of GCRL

Our Finding: Inductive bias in $V(s, g)$ mitigates overgeneralization!



- GCIVL + IQE also exhibits accurate generalization.
- QRL + MLP leads to training failure.

$\Rightarrow$ Overgeneralization depends on the architecture of $V(s, g)$.

Nevertheless, neither GCIVL + IQE nor QRL + IQE are still far from a satisfactory solution, due to their large performance variance.

Key Findings

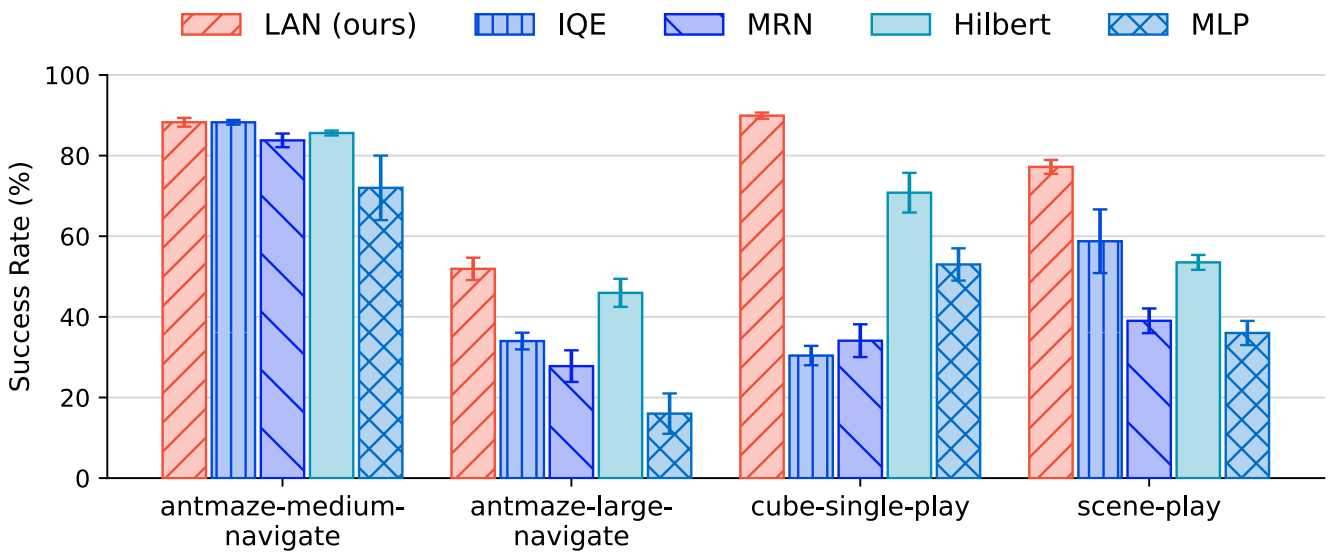
Finding 1: Structural Inductive Bias

While the standard MLP-parameterized value function suffers from value overgeneralization, value functions with **structural inductive bias** can mitigate overgeneralization.

Finding 2: Limitations of Quasimetric Method

(i) The QRL objective induces training instability in some tasks, highlighting **the necessity of TD learning** for robust performance.
(ii) While GCIVL + IQE improves upon GCIVL + MLP by mitigating overgeneralization, **the IQE itself can degrade performance** in some tasks.

LAN: Latent Alignment Network



We introduce **Latent Alignment Network (LAN)**, a new value function architecture that provides effective inductive bias for value generalization:

$$V(s, g) = -\|\varphi_S(s) - \varphi_G(g)\|_2, \text{ where } \varphi_S : \mathcal{S} \to \mathbb{R}^d, \varphi_G : \mathcal{G} \to \mathbb{R}^d$$

- Represents goal-conditioned value as negative Euclidean distance in learned latent space.
- Empirically outperforms existing value function architectures.

LAVL: Latent-Aligned Value Learning

We propose **Latent-Aligned Value Learning (LAVL)**, an offline GCRL algorithm leveraging LAN-based value learning and hierarchical policy.

1. Value Learning with LAN

- Train LAN value via expectile regression (Kostrikov et al., 2022):

$$\mathcal{L}_{\text{TD}}(V) = \mathbb{E}_{\substack{(s,s')\in\mathcal{D}, \\ g\sim p^{\mathcal{D}}_{\text{mixed}}(\cdot|s)}} [\ell_2^\kappa(r(s,g)+\gamma\widetilde{V}(s',g)-V(s,g))]$$

- Local continuity regularization to stabilize training:

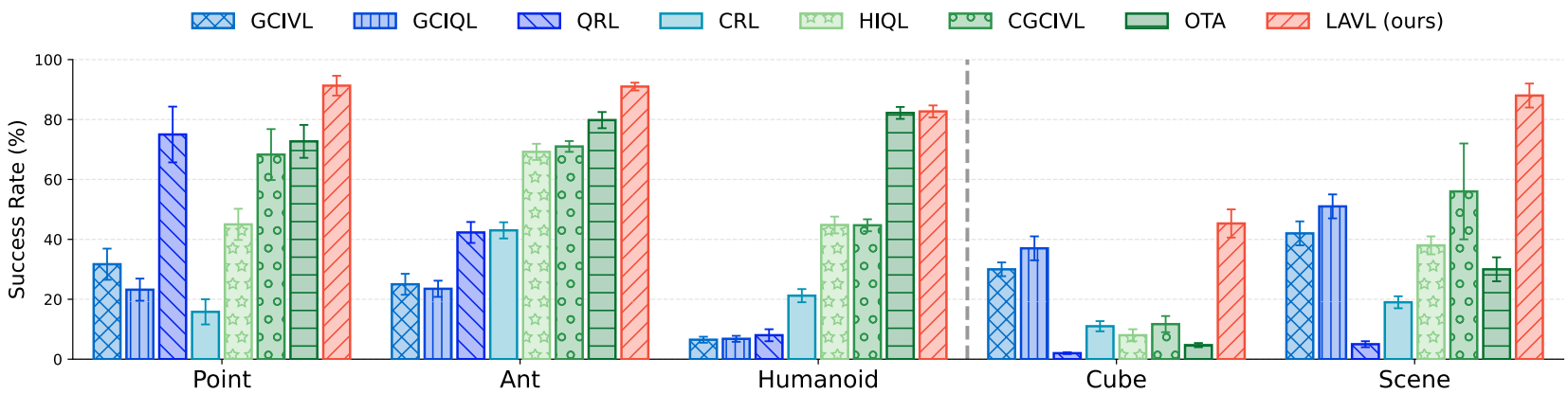
$$\mathcal{L}_{\text{Reg}}(V) = \mathbb{E}_{\substack{(s,s')\in\mathcal{D}, \\ g\sim p^{\mathcal{D}}_{\text{rand}}(\cdot|s)}} \left[\left(\left(V(s,g)-V(s',g) \right)^2 - \delta^2 \right)_+ \right].$$

- Final loss: $\mathcal{L}(V) = \mathcal{L}_{\text{TD}}(V) + w_c \mathcal{L}_{\text{Reg}}(V)$ with $w_c \geq 0$.

2. Hierarchical Policy Extraction

- Adopt the hierarchical policy extraction of HIQL (Park et al., 2023), where high / low level policies are trained via AWR (Peng et al., 2019).

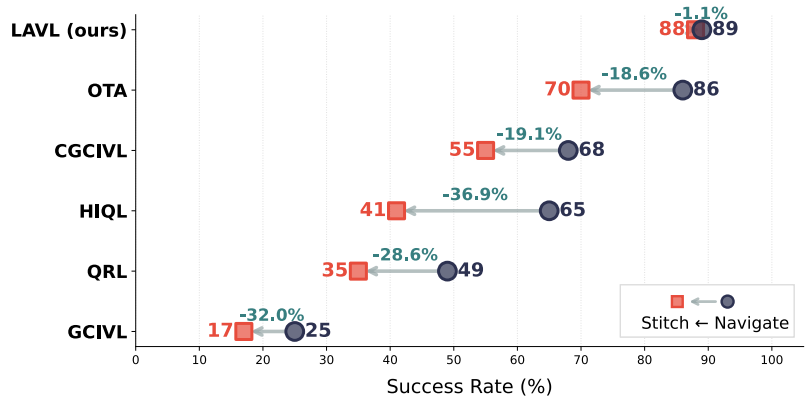
Experiments



- Evaluation on 22 OGBench (Park et al., 2025) datasets, encompassing robotic locomotion and manipulation tasks.
- LAVL achieves the best performance in **20 out of 22 tasks**, consistently outperforming strong baselines.
- Baselines often perform well only on either locomotion or manipulation tasks, but LAVL shows consistent performance across both domains.

Experiments

Algorithm	Medium	Large	Giant	Rel. Drop (%) ↓
LAVL (ours)	**94**	**85**	**85**	**9.6%**
OTA	89	78	68	23.1%
CGCIVL	84	66	35	58.3%
QRL	58	45	22	61.7%
HIQL	87	51	21	75.4%
GCIVL	48	16	0	100.0%



- **Long Horizon.** While baselines suffer performance degradation as task horizon increases (from `medium` to `giant`), LAVL exhibits remarkable stability.
- **Stitching Trajectories.** Baseline methods exhibit substantial degradation when trajectory stitching is required. In contrast, LAVL can stitch trajectories with minimal performance drop.

Summary

- **Our Finding.** Erroneous generalization in goal-conditioned value is a fundamental bottleneck, and inductive bias is crucial for addressing the bottleneck.
- **Our Solution.** We propose LAVL, an offline GCRL algorithm that integrates latent-representation-based value generalization with hierarchical policy.
- **Experiments.** LAVL consistently outperforms baselines on OGBench datasets. Notably, LAVL exhibits strong performance in long-horizon tasks and trajectory stitching datasets.