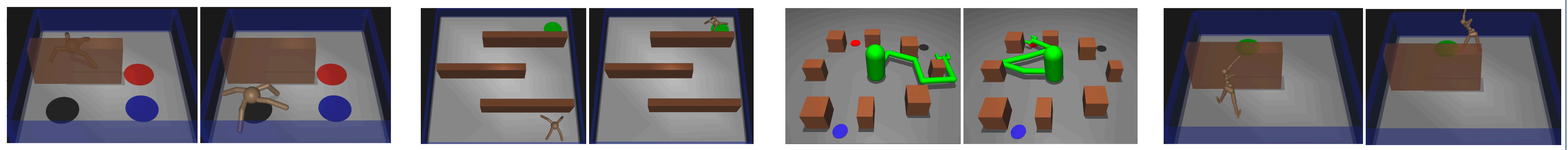
Universal Planning Networks

Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, Chelsea Finn

Code: https://github.com/aravind0706/upn

Problem of Interest

Visual continuous control tasks specified via **target (goal) image** and **no extrinsic rewards**



Can we learn representations via imitation learning that support goal-directed planning for such tasks?

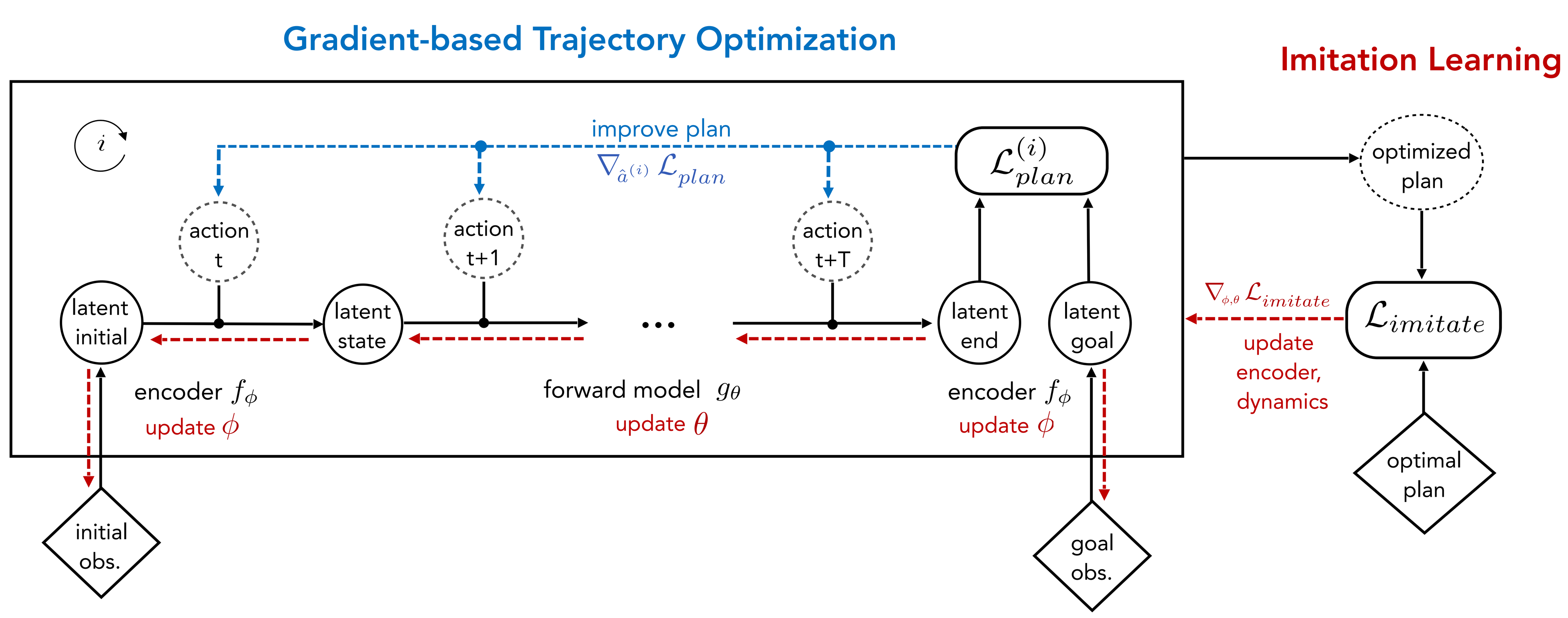
Proposed Approach

A model-based goal-conditioned policy architecture that

- ✤ performs differentiable planning in a latent space jointly learned with forward dynamics
- ✤ is trained end-to-end to learn representations with which gradient-based trajectory optimization leads to successful visuomotor policy imitation.

Thus, the learned underlying state representations are directly optimized to encode task-related information about the visual environment that supports goal-directed planning.

Universal Planning Networks (UPN)



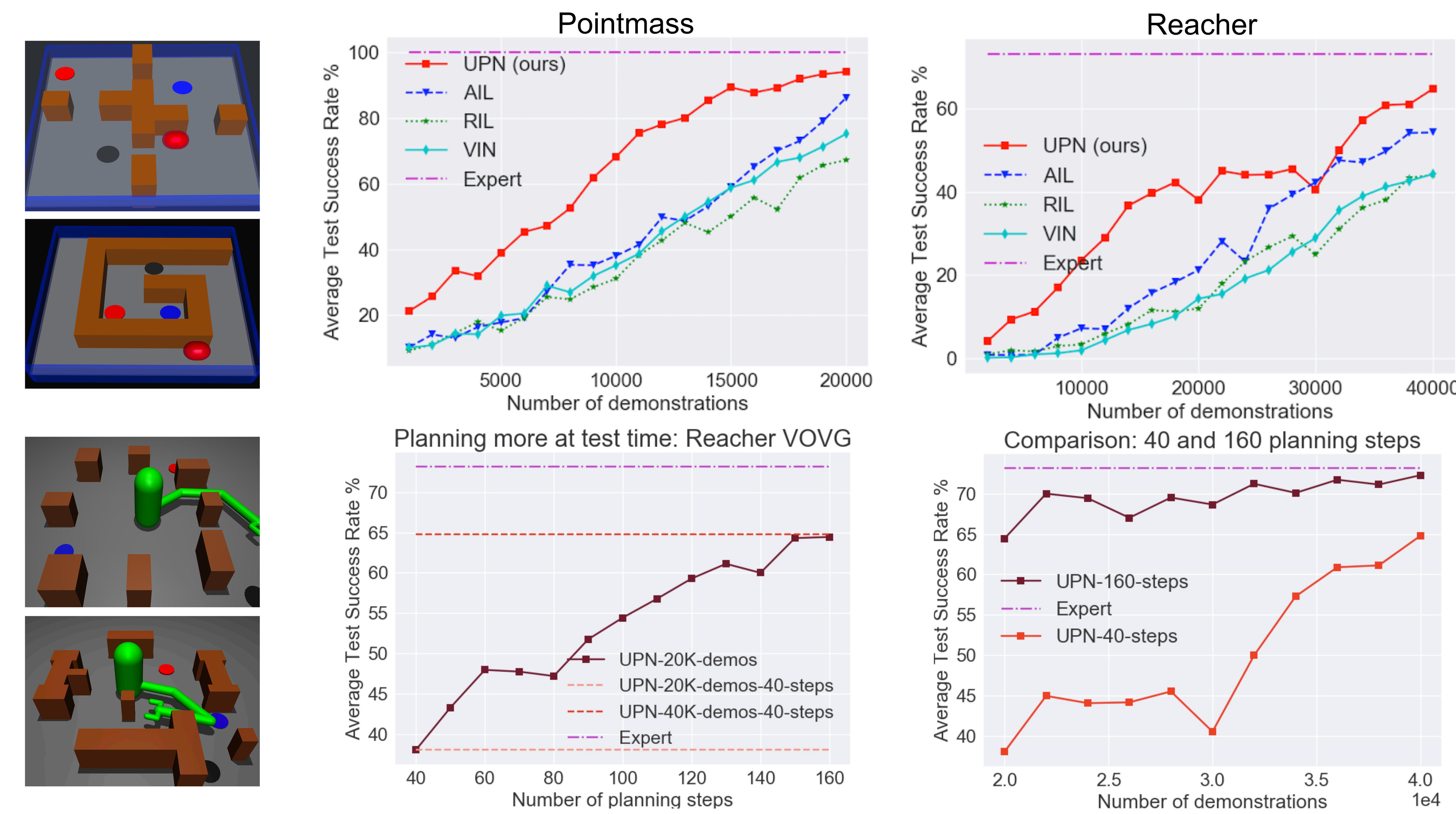
- ✤ The encoder is a CNN and the dynamics model is an RNN.
- ✤ The gradient for the parameter update is **back-propagated through the iterative planning procedure.**
- ✤ We found that number of inner planning steps can be upwards of 100.

Motivation for Experiments

Our experiments were designed to answer the following key questions:

1. Can UPNs learn effective visual imitation policies that can generalize to unseen goals at test time?
2. Are these policies better (in terms of sample efficiency) than conventional methods?
3. Can the learned gradient-based planning be exploited at test-time?
4. Can we use the learned UPN representations to specify rewards for new goal-conditioned visuomotor tasks?

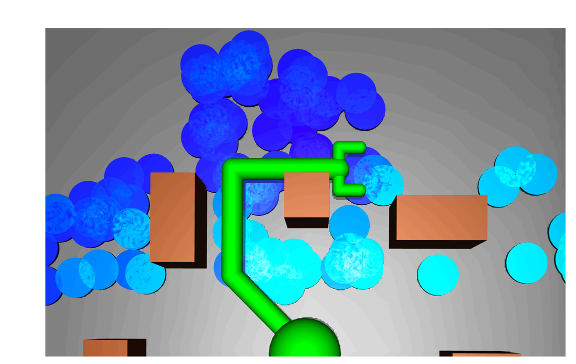
Imitation Learning Results & the Benefit of Extra Planning at Test-time

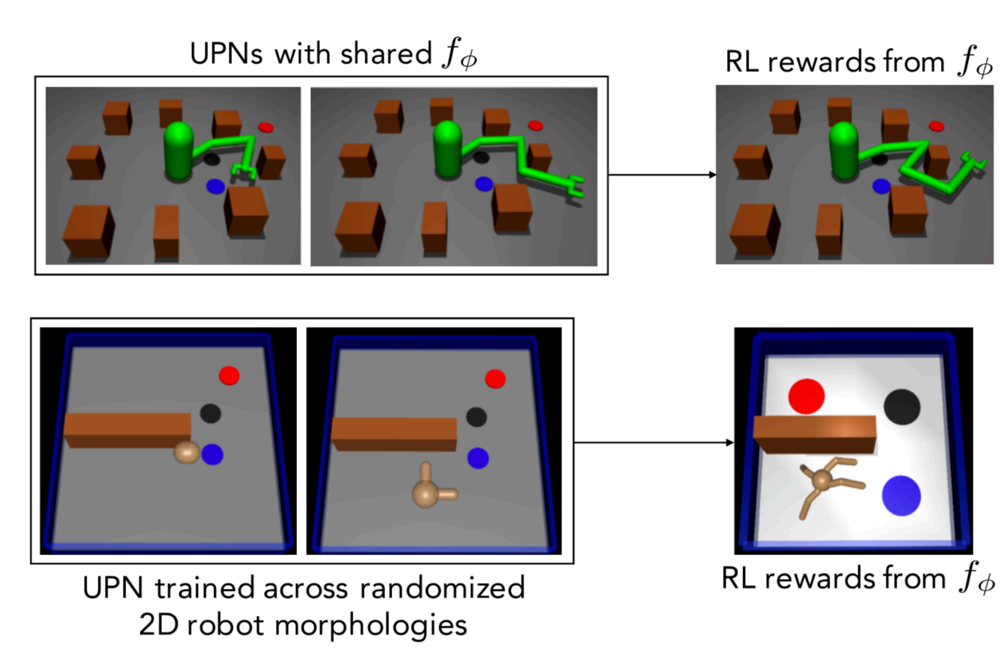


- ✤ UPN is more sample-efficient than conventional methods for effective visual imitation.
- ✤ Using more planning iterations at test time helps, even if the UPN was trained with fewer planning iterations.

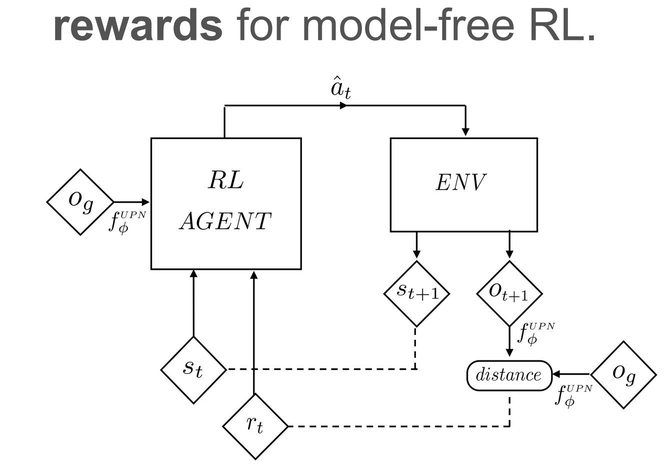
Beyond Imitation: Transferring to significantly different tasks using RL

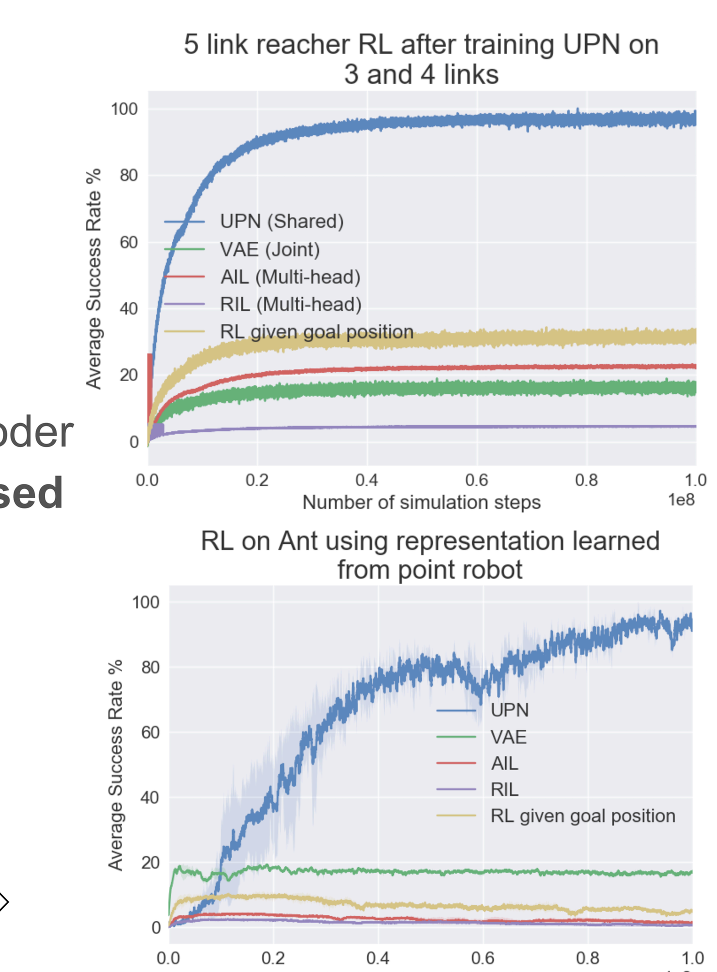
✤ By relying on minimizing the latent distance-to-goal throughout the planning process, the UPN **learns an optimizable latent distance metric** that encodes task-specific semantics.





✤ We can use the acquired encoder to specify **latent distance-based rewards** for model-free RL.

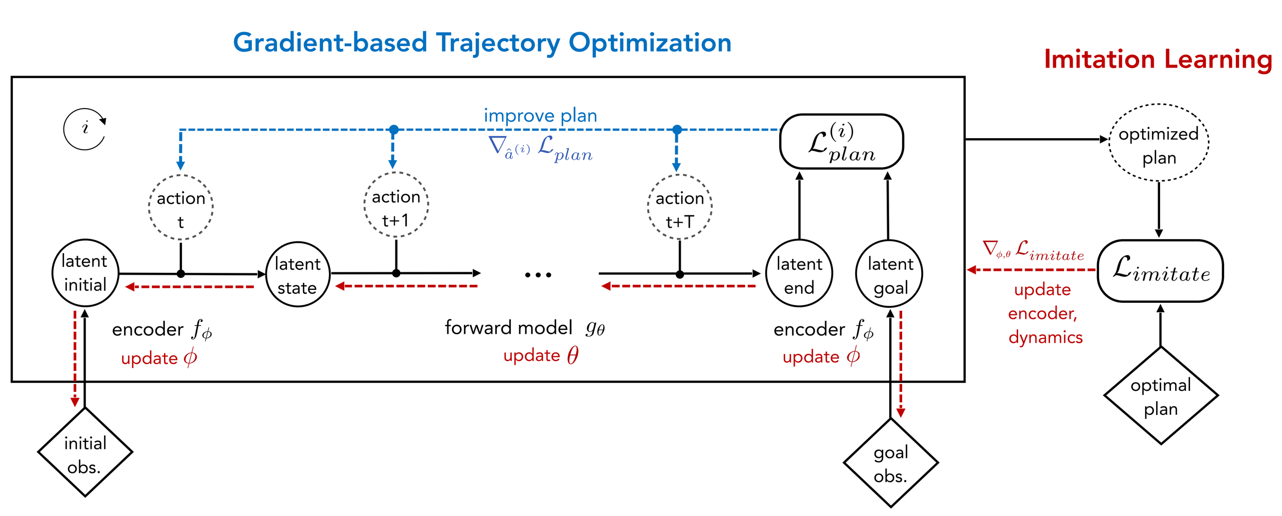




Takeaways

- ✤ UPNs learn effective visual goal-directed policies efficiently.
- ✤ Learned latent representations can be leveraged to **transfer task-related semantics** through goal-conditioned reward functions.
- ✤ The learned planner **benefits from more updates at test-time**, suggesting that acquired representations are general

Supervised objectives can be useful for improving reinforcement learning agents.



Gradient-based Trajectory Optimization
Imitation Learning
improve plan
$\nabla_{\hat{a}^{(i)}} \mathcal{L}_{plan}$
$\mathcal{L}^{(i)}_{plan}$
i
action t
action t+1
action t+T
latent initial
latent state
...
latent end
latent goal
encoder f_ϕ
update ϕ
forward model g_θ
update θ
encoder f_ϕ
update ϕ
initial obs.
goal obs.
optimized plan
$\nabla_{\phi,\theta} \mathcal{L}_{imitate}$
$\mathcal{L}_{imitate}$
update encoder, dynamics
optimal plan