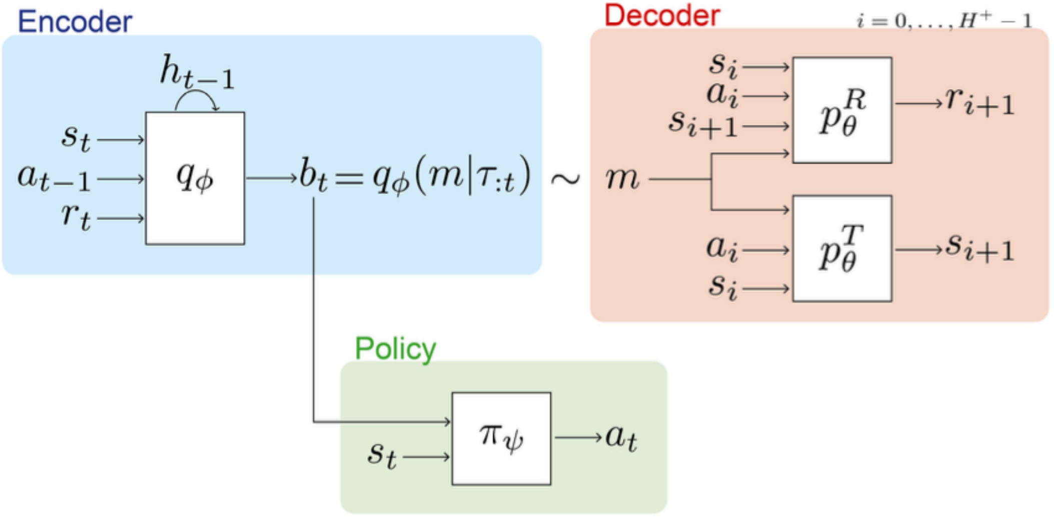
Zero-Shot Meta-RL: VariBAD [Zintgraf et al. 2020]

- Learn unifying Reward and Transition distribution q_ϕ of unknown MDPs M_i using latent variable m_i
  - MDPs Share some structure
- Encoder encodes history information $\tau_{:t}^{(i)}$ gathered online during learning in M_i
- Decoder decodes the whole trajectory including a future state



[Zintgraf et al. 2020] - Figure 2

Zero-Shot Meta-RL: VariBAD [Zintgraf et al. 2020]

- The Policy then uses the belief b_t about the MDPs in its decision making
  - Since there is now information about the MDPs this should give a good trade-off between exploration and exploitation in all of the MDPs

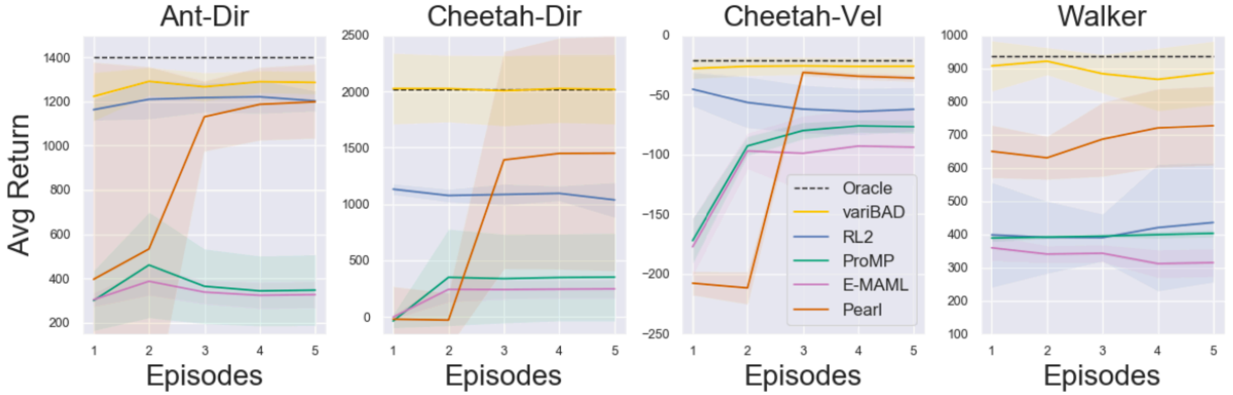


Figure 4: Average test performance for the first 5 rollouts of MuJoCo environments (using 5 seeds).