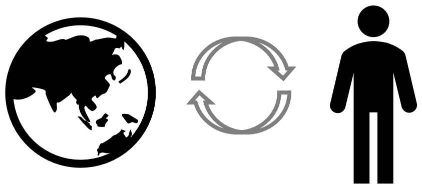
Learning Versatile Skills with Curriculum Masking

Yao Tang[1]*, Zhihui Xie[1]*, Zichuan Lin[2], Deheng Ye[2], Shuai Li[1]

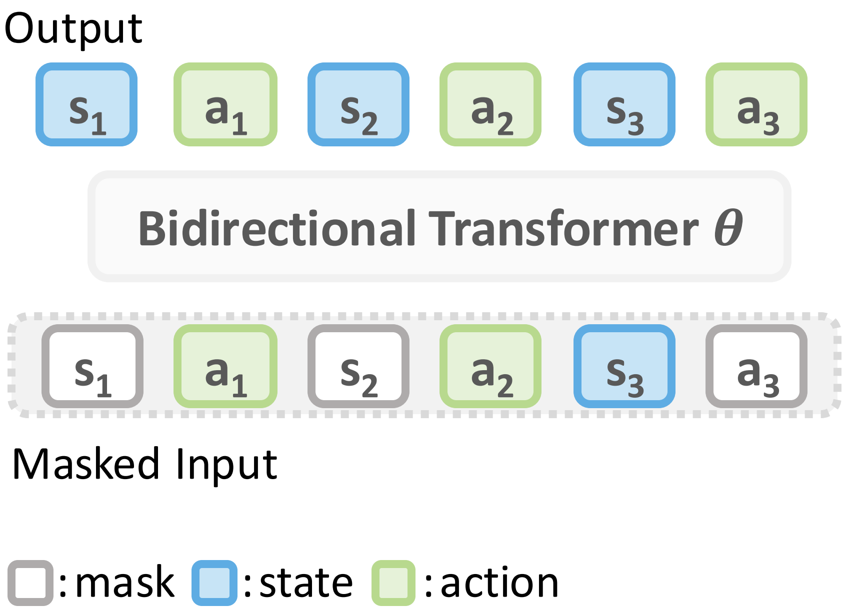
[1]Shanghai Jiao Tong University [2]Tencent AI Lab

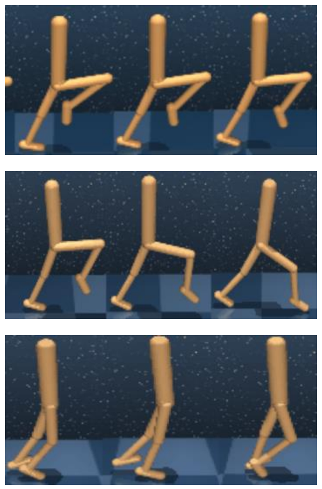
yaotang923@gmail.com

Unsupervised RL Pretraining



$(s_1, a_1, s_2, a_2, s_3, a_3)$





Ingredient 1

Reward-free offline data

Ingredient 2

Masked prediction[1]

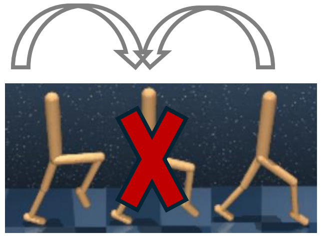
A unified model learns versatile skills[1,2]

[1] Liu et al., 2022, Masked Autoencoding for Scalable and Generalizable Decision Making

[2] Sun et al. , 2023, SMART: Self-supervised Multi-task pretrAining with contRol Transformers

Masked Prediction on Decision-making Data

- A mask scheme = A reusable skill (s_1,[MASK],s_2,a_2,[MASK], a_3)
- Random masking[1]?



s_1 a_1 s_2 a_2 s_3 a_3

"Heavy Information redundancy[1]"

"Interleaved modality"

Our research question:

How to design & arrange mask schemes for decision-making data?

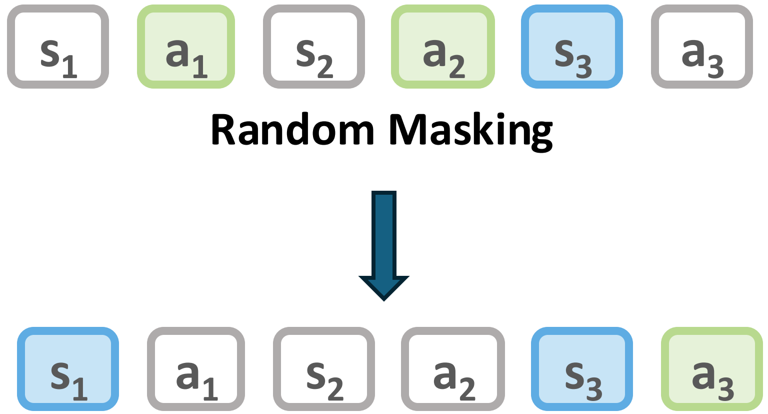
[1] Liu et al., 2022, Masked Autoencoding for Scalable and Generalizable Decision Making

Curriculum Masking

How to **design** & **arrange** mask schemes for decision-making data?

- Main intuition: humans organize knowledge in a curriculum, from easy to hard

- **Block-wise masking**: a semantic entity of skill

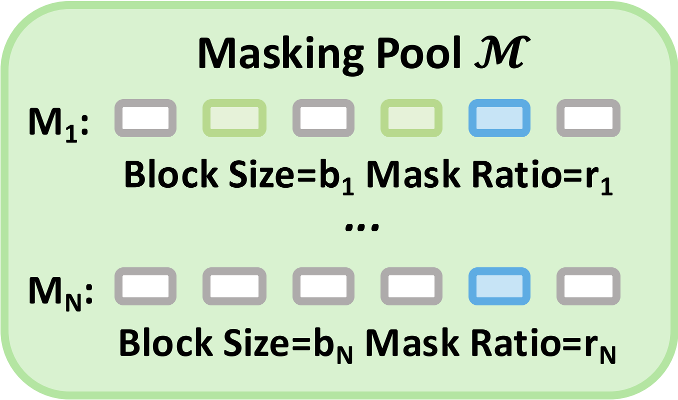


Block-wise Masking (block size=3)

- Small block size & mask ratio: local dynamics
- Large block size & mask ratio: global dependency

Curriculum Learning

- Core of Curriculum Masking: dynamically adjust mask schemes based on the learning progress



- **Evaluate learning progress**: target loss decrease[1]

$$r = f_{\text{scale}}(\mathcal{L}_{\text{target}}(\theta) - \mathcal{L}_{\text{target}}(\theta'))$$

- **Select masking schemes based on learning progress**: multi-armed bandit algorithm EXP3[2]

$$\pi_{\mathbf{w}}(i) = (1-\epsilon)\frac{w_i}{\sum_{j=1}^{K} w_j} + \frac{\epsilon}{K} \quad i = 1, \ldots, K$$

[1] Graves et al., 2017, Automated curriculum learning for neural networks

[2] Auer et al., 2002, The nonstochastic multiarmed bandit problem

Downstream Performance

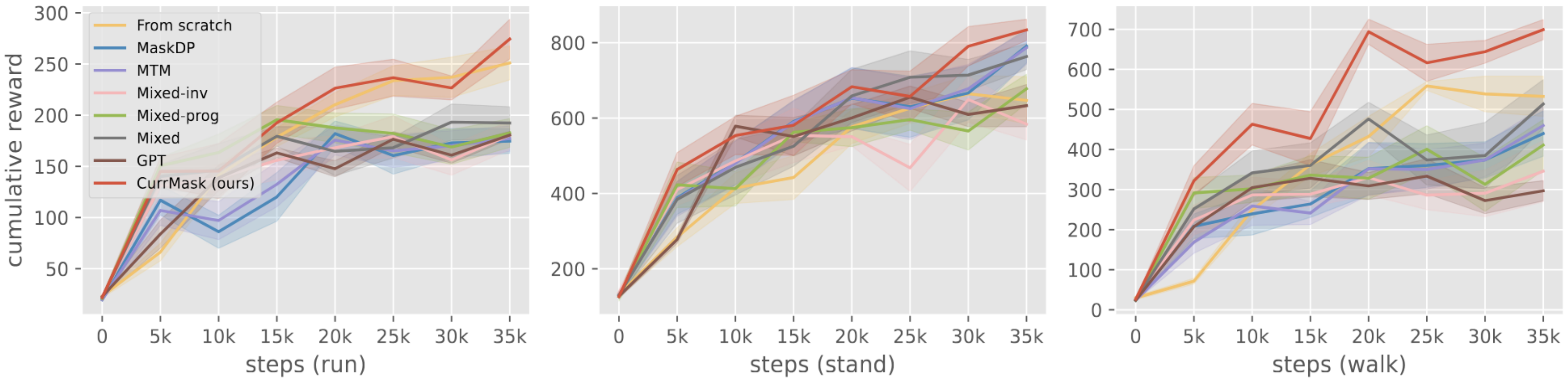
- Skill Prompting

Reward ↑	walker_s	walker_w	walker_r	quad_w	quad_r	jaco_bl	jaco_br	jaco_tl	jaco_tr	Average
MaskDP	103.2 ± 2.6	58.4 ± 2.3	29.3 ± 1.4	36.6 ± 2.2	45.1 ± 2.4	58.1 ± 4.4	58.4 ± 3.0	56.9 ± 3.9	64.0 ± 3.3	56.7
MTM	107.1 ± 2.8	58.8 ± 2.7	27.3 ± 1.4	37.1 ± 2.3	42.8 ± 2.7	72.0 ± 4.5	71.8 ± 3.9	72.5 ± 5.1	77.6 ± 3.5	63.0
Mixed-inv	103.3 ± 3.1	59.5 ± 3.0	23.8 ± 1.2	**45.1** ± 3.0	43.2 ± 2.9	51.8 ± 3.3	53.0 ± 2.8	56.8 ± 3.6	59.7 ± 4.8	55.1
Mixed-prog	103.5 ± 2.4	55.0 ± 2.8	25.8 ± 1.2	40.5 ± 1.8	45.6 ± 2.2	85.3 ± 5.5	85.4 ± 3.7	84.2 ± 4.8	88.5 ± 3.7	68.2
Mixed	110.8 ± 2.2	54.2 ± 2.0	30.5 ± 1.2	43.3 ± 2.7	**51.3** ± 2.8	66.0 ± 6.4	61.6 ± 3.7	62.3 ± 3.6	66.5 ± 4.0	60.7
GPT	101.8 ± 2.9	34.6 ± 1.3	21.6 ± 1.0	41.9 ± 2.9	48.8 ± 3.2	86.1 ± 5.7	83.1 ± 2.7	83.9 ± 5.1	85.7 ± 3.0	65.3
CurrMask	**111.2** ± 2.4	**79.9** ± 1.2	**38.9** ± 1.9	38.0 ± 2.2	51.0 ± 3.4	**88.4** ± 5.1	**88.5** ± 3.6	**86.0** ± 4.3	**92.9** ± 3.5	**75.0**

- Goal-conditioned Planning

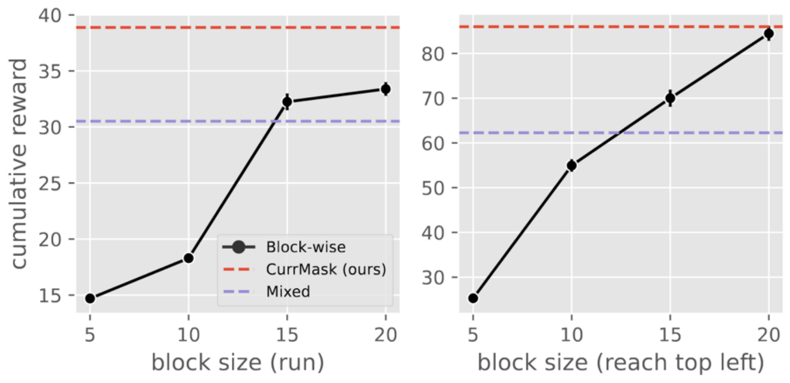
Distance ↓	walker_s	walker_w	walker_r	quad_w	quad_r	jaco_bl	jaco_br	jaco_tl	jaco_tr	Average
MaskDP	4.85 ± 0.48	10.10 ± 0.27	15.52 ± 0.39	20.71 ± 0.69	21.62 ± 0.79	1.42 ± 0.05	1.42 ± 0.05	1.39 ± 0.06	1.40 ± 0.06	8.71
MTM	6.05 ± 0.61	12.20 ± 0.41	17.92 ± 0.55	23.93 ± 0.70	25.09 ± 0.80	2.38 ± 0.08	2.42 ± 0.10	2.35 ± 0.07	2.30 ± 0.10	10.59
Mixed-inv	5.32 ± 0.53	11.25 ± 0.31	16.51 ± 0.51	22.63 ± 0.74	23.31 ± 0.77	1.55 ± 0.06	1.53 ± 0.05	1.57 ± 0.07	1.53 ± 0.08	9.47
Mixed-prog	4.96 ± 0.48	10.18 ± 0.28	15.77 ± 0.48	23.49 ± 0.72	24.28 ± 0.86	1.46 ± 0.04	1.44 ± 0.04	1.44 ± 0.05	1.44 ± 0.09	9.38
Mixed	**4.83** ± 0.47	10.15 ± 0.28	15.47 ± 0.46	20.67 ± 0.73	21.66 ± 0.75	1.47 ± 0.06	1.47 ± 0.04	1.43 ± 0.06	1.44 ± 0.08	8.73
Goal-GPT	7.47 ± 0.74	15.15 ± 0.41	21.04 ± 0.60	27.36 ± 0.77	28.76 ± 0.90	3.34 ± 0.10	3.58 ± 0.11	3.26 ± 0.15	3.50 ± 0.11	12.61
CurrMask	4.85 ± 0.47	**9.90** ± 0.27	**15.31** ± 0.49	**20.47** ± 0.71	**21.39** ± 0.67	**1.39** ± 0.05	**1.38** ± 0.04	**1.33** ± 0.05	**1.34** ± 0.07	**8.60**

- Offline RL

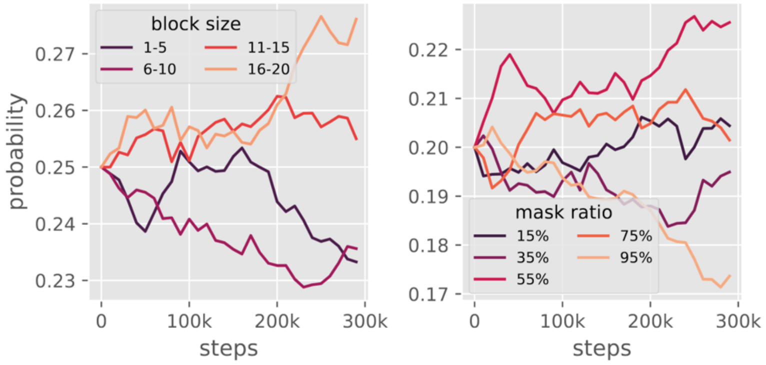


CurrMask consistently outperforms other baselines on various downstream tasks

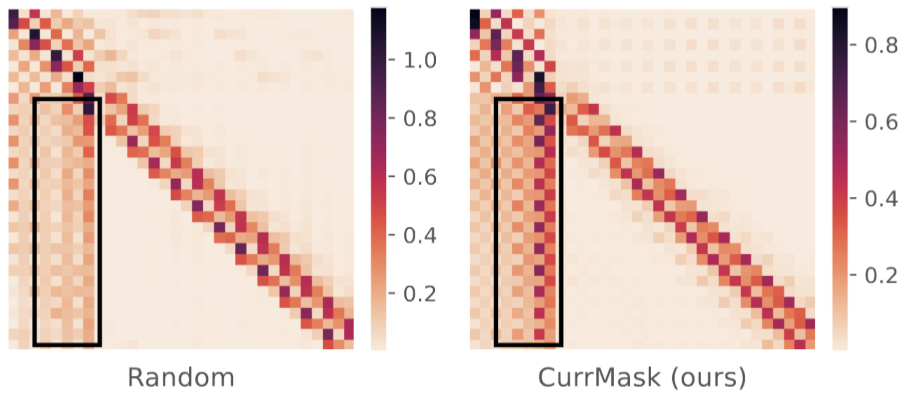
Analysis



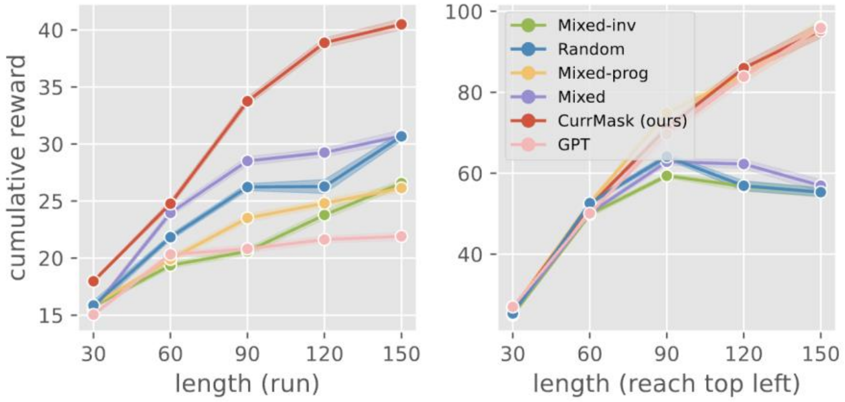
Impact of Block-wise Masking



Impact of Masking Curricula



Attention Maps



Skill Prompting reward v.s. rollout length

Summary

- Curriculum Masking for unsupervised RL pretraining
  - A unified model to learn **versatile** skills
  - **Adaptivity** in adjusting learning strategy
  - Superior ability to extract **local dynamics** & **global dependencies**
- Limitations
  - A training time (wall clock time) overhead of 4.7%
  - Advantages could be affected by the underlying structure of the environment