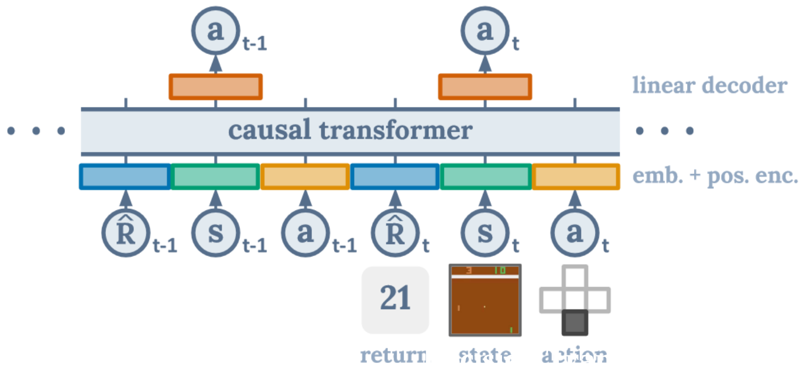
Decomposed Prompt Decision Transformer for Efficient Unseen Task Generalization

Hongling Zheng[1], Li Shen[2†], Yong Luo[1†], Tongliang Liu[3], Jialie Shen[4], Dacheng Tao[5]

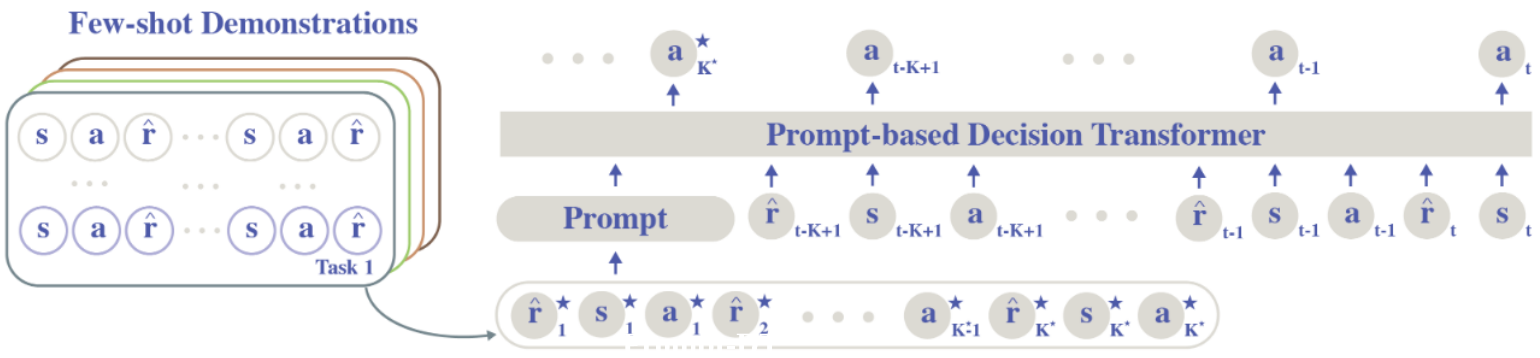
[1]*Wuhan University* [2]*Shenzhen Campus of Sun Yat-sen University* [3]*The University of Sydney*
[4]*City, University of London* [5]*Nanyang Technological University*

NeurIPS 2024, the Thirty-Eighth Annual Conference on Neural Information Processing Systems

Background: Sequenced-based Offline RL



Decision Transformer

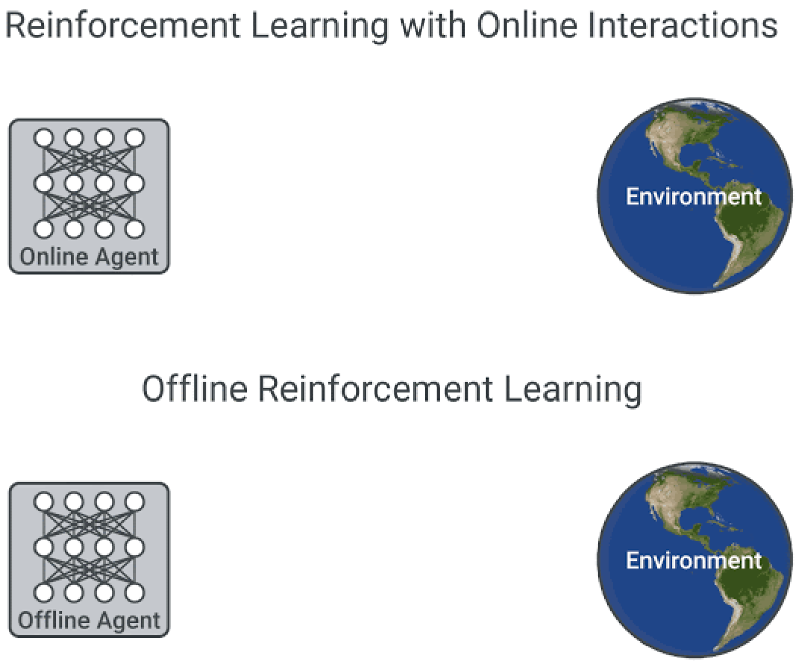


Prompt-based Decision Transformer

Sequenced-based offline RL algorithms abandon the traditional dynamic programming approach in offline RL and adopt an autoregressive paradigm.

- Decision Transformer models trajectories using tuples of returns, states, and actions collected at different time steps. Here, returns denote the cumulative reward from the current time step until the end of the episode.
- Prompt-DT formalizes offline RL as a few-shot policy generalization problem. It is trained on a set of tasks with prompts and offline data, enabling it to generalize to new tasks.

Motivation



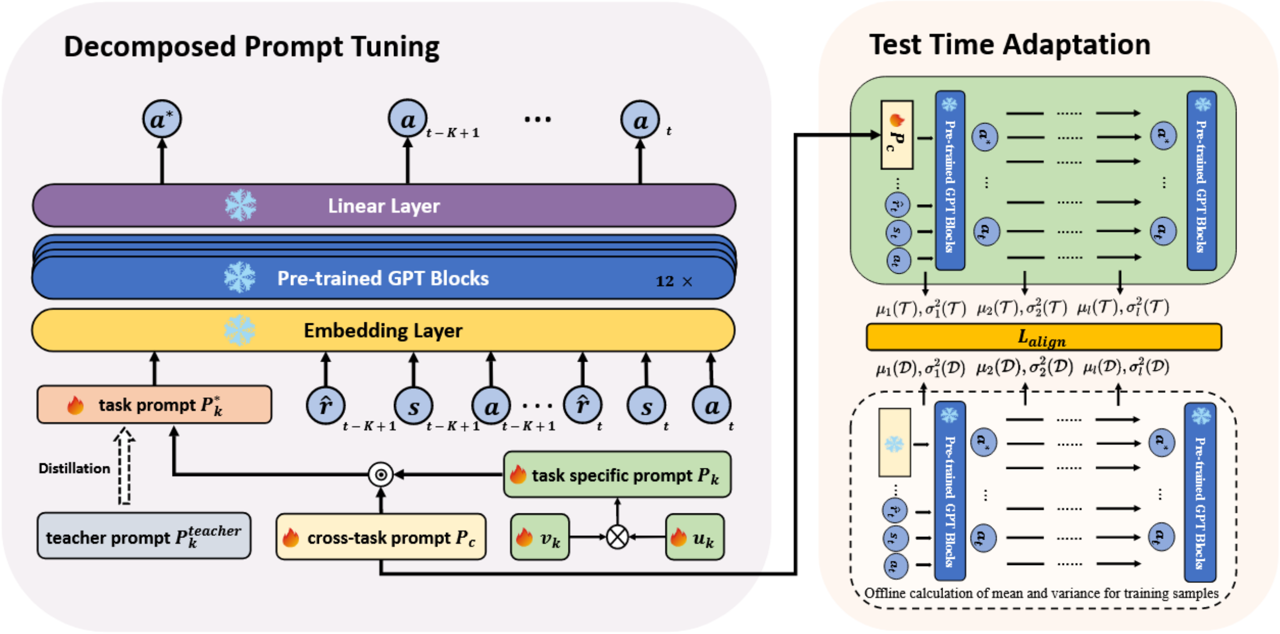
- Traditional offline RL algorithms rely heavily on historical trajectory data, and models may overestimate values for unseen actions or states, resulting in suboptimal policies.
- While sequence-based offline RL methods like fine-tuning work well in specific scenarios, they often require task-specific data, limiting their applicability, especially when target task data is unavailable.
- How can we better explore relationships between tasks to extract cross-task prompts and generalize to new downstream tasks?

Contribution

- We designed DPDT (Decomposed Prompt Decision Transformer) for knowledge extraction and zero-shot generalization.

- By leveraging a pre-trained language model, DPDT decomposes multi-task prompts into cross-task and task-specific prompts through distillation.

- During testing, DPDT uses test-time adaptation (TTA) to optimize prompts by aligning them with hidden layer features of unlabeled test and training data.



Method: Decomposed Prompt Tuning

- **Initialization:**

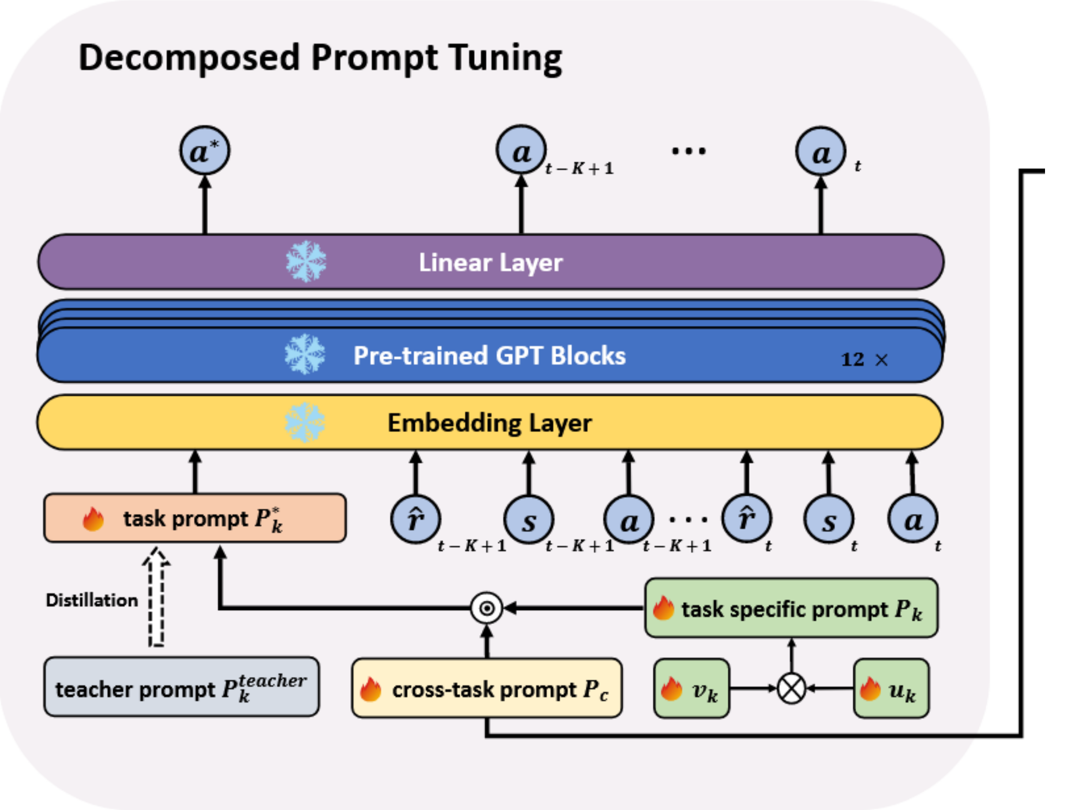
DPDT is initialized using the pre-trained language model GPT2-small.

- **Prompt Decomposition:**

Given a set of training tasks $S = (S_1, S_2, \dots, S_n)$, the cross-task prompt P_c is designed to capture shared knowledge from S, while the task-specific prompt P_k allows each task to retain its unique knowledge. To reduce computational complexity in the implementation, P_k is further decomposed into two low-rank vectors $v_k \in l * r, \mathrm{u}_k \in r * s$.

$$P_k^* = P_c \circ P_k = P_c \circ (v_k \otimes u_k)$$

$$\mathcal{L}_{MSE} = (a - \mathcal{M}(P_k^\star, \tau))^2$$



Method: Decomposed Prompt Tuning

- **Prompt Distillation:**

Due to the lack of explicit constraints in the specific implementation process, directly implementing prompt decomposition on the multitask dataset S may lead to an overlap in the information learned by P_c and P_k, potentially undermining their ability to capture distinct intended details.

$$\mathcal{L}_{dis} = \sum_{k \in |\mathcal{S}|} |p_k^{teacher} - p_k^{\star}|^2$$

$$\mathcal{L}_{Total} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{dis}$$

Algorithm 1 Decomposed Prompt Tuning

Input: Training task set S, Offline datasets $\mathcal{D}_{\mathcal{M}}$, Batch size M, Learning rate α, training iterations N, teacher task prompts $p_k^{teacher}$.
Initialize: Initialize a 12-layer, 12-head DPDT $\mathcal{M}$ using GPT2-SMALL, randomly initialize cross-task prompts P_c and low-rank vectors v_k, u_k.

```
for t = 1 to N do
    for k in S do
        Select a trajectory τ that contains M samples in task k.
        Calculate P_k^* by Equation 3.
        Calculate L_MSE and L_dis according to Equations 4 and 5.
        Computed loss function by Equation 6.
        θ ← θ − α∇_θ L_Total.
    end for
end for
```

Method: Test Time Adaptation

- **Calculating the aligned token mean and variance**

$$\mu_l(\mathcal{T}) = \frac{1}{|X|}\sum_{i=1}^{|X|} H_{l,i}, \quad \sigma_l^2(\mathcal{T}) = \frac{1}{|X|}\sum_{i=1}^{|X|}[H_{l,i} - \mu_l(\mathcal{T})]^2$$

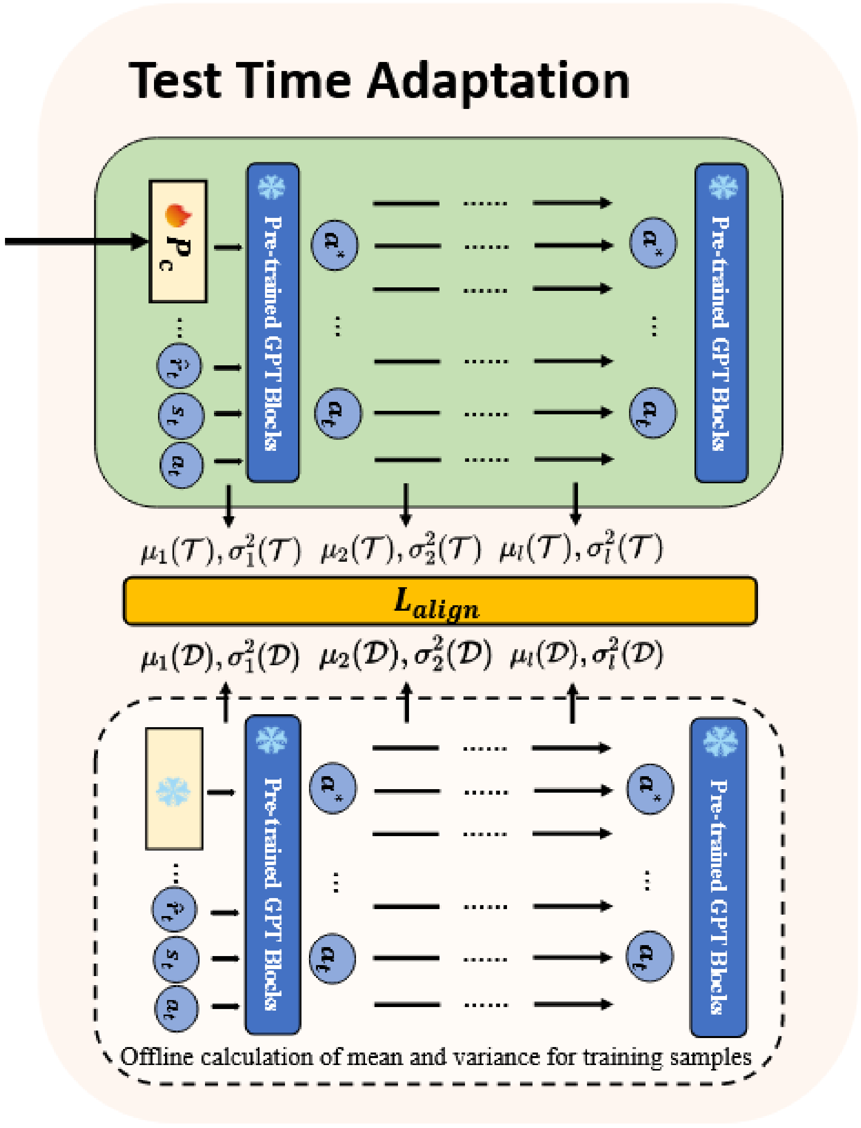
- **Calculating the alignment loss function**

$$L_{\text{align}} = \frac{1}{L}\sum_{l=1}^{L}\left(\|\mu_l(\mathcal{T}) - \mu_l(\mathcal{D})\|_1 + \|\sigma_l^2(\mathcal{T}) - \sigma_l^2(\mathcal{D})\|_1\right)$$

Algorithm 2 Test Time Adaptation

Input: Test samples set X, Cross-task prompts P_c, $\mu_l(\mathcal{D})$, $\sigma_l^2(\mathcal{D})$, The number of layers L.

```
for l = 1 to L do
    for i in X do
        Calculate H_{l,i} obtained by inputting the concatenation of P_c and i into DPDT.
    end for
end for
for l = 1 to L do
    Compute μ_l(T) and σ_l^2(T) by Equation 7.
end for
Compute token distribution alignment loss by Equation 8.
Optimize L_align to update P_c.
```



Experiment: Main Results

- **Table 1: Results for Meta-RL control tasks (zero-shot scenarios).**

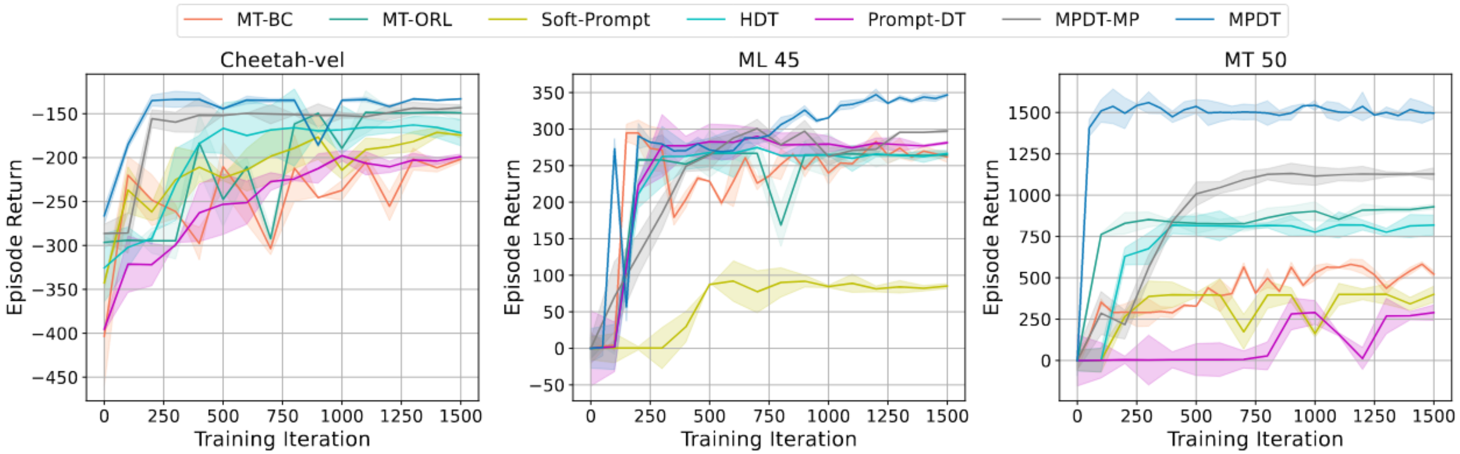
	MT-BC [44]	MT-ORL [5]	Soft-Prompt [45]	HDT [17]	Prompt-DT [6]	DPDT-WP	DPDT
Trainable Params	125.5M	125.5M	3.94M	12.94M	125.5M	1.42M	1.42M
Percentage	100%	100%	3%	10.31%	100%	1.14%	1.14%
Cheetah-dir	$-24.71_{\pm 12.04}$	$-86.92_{\pm 15.51}$	$-4.21_{\pm 5.51}$	$-45.32_{\pm 13.22}$	$-7.92_{\pm 2.97}$	$11.73_{\pm 12.8}$	$\mathbf{50.32_{\pm 11.47}}$
Cheetah-vel	$-201.66_{\pm 30.27}$	$-148.24_{\pm 22.18}$	$-171.23_{\pm 20.58}$	$-162.75_{\pm 20.50}$	$-192.38_{\pm 11.80}$	$-143.14_{\pm 21.40}$	$\mathbf{-139.88_{\pm 19.65}}$
Ant-dir	$131.89_{\pm 12.96}$	$109.21_{\pm 9.66}$	$119.45_{\pm 14.2}$	$115.43_{\pm 10.22}$	$\mathbf{123.46_{\pm 10.70}}$	$101.49_{\pm 17.74}$	$121.84_{\pm 8.01}$
MW ML10	$256.77_{\pm 11.93}$	$343.16_{\pm 9.40}$	$246.42_{\pm 24.60}$	$292.14_{\pm 8.21}$	$317.31_{\pm 14.98}$	$204.88_{\pm 28.96}$	$\mathbf{371.01_{\pm 9.41}}$
MW ML45	$287.37_{\pm 11.38}$	$266.744_{\pm 25.81}$	$91.97_{\pm 14.11}$	$274.88_{\pm 19.74}$	$294.55_{\pm 8.71}$	$300.71_{\pm 15.74}$	$\mathbf{347.21_{\pm 11.52}}$
MW MT 10	$547.83_{\pm 11.04}$	$1064.58_{\pm 21.70}$	$201.23_{\pm 7.11}$	$964.57_{\pm 15.34}$	$1087.54_{\pm 17.09}$	$1015.91_{\pm 0.74}$	$\mathbf{1317.52_{\pm 8.22}}$
MW MT 50	$582.80_{\pm 13.48}$	$929.74_{\pm 22.81}$	$400.71_{\pm 26.40}$	$820.45_{\pm 27.19}$	$994.63_{\pm 5.99}$	$1131.01_{\pm 1.17}$	$\mathbf{1559.94_{\pm 2.49}}$
Average	225.76	354.04	130.62	309.79	373.88	374.66	**518.28**

- **Table 2: Results for Meta-RL control tasks (few-shot scenarios).**

	MT-ORL [5]	Soft-Prompt [45]	HDT [17]	Prompt-DT [6]	DPDT-WP	DPDT	DPDT-F
Trainable Params	125.5M	3.94 M	12.94 M	125.5M	1.42M	1.42M	125.5M
Percentage	100%	3%	10.31%	100%	1.14%	1.14%	100%
Cheetah-dir	$-46.22_{\pm 3.44}$	$940.24_{\pm 1.08}$	$875.23_{\pm 4.24}$	$934.78_{\pm 5.33}$	$946.81_{\pm 17.24}$	$\mathbf{955.17_{\pm 8.03}}$	$1037.85_{\pm 5.98}$
Cheetah-vel	$-146.64_{\pm 2.12}$	$-41.81_{\pm 2.10}$	$-63.81_{\pm 6.30}$	$-37.80_{\pm 2.09}$	$-48.07_{\pm 1.85}$	$\mathbf{-30.73_{\pm 1.88}}$	$-29.85_{\pm 9.46}$
Ant-dir	$110.51_{\pm 2.2}$	$379.01_{\pm 1.75}$	$361.49_{\pm 5.63}$	$\mathbf{411.96_{\pm 9.28}}$	$308.10_{\pm 5.22}$	$384.29_{\pm 10.91}$	$400.01_{\pm 9.79}$
MW ML10	$421.22_{\pm 9.21}$	$379.82_{\pm 14.76}$	$467.81_{\pm 3.07}$	$315.07_{\pm 6.17}$	$485.27_{\pm 19.31}$	$\mathbf{535.52_{\pm 17.39}}$	$670.24_{\pm 3.88}$
MW ML45	$264.14_{\pm 9.67}$	$448.72_{\pm 11.38}$	$477.19_{\pm 2.16}$	$473.34_{\pm 4.12}$	$519.28_{\pm 7.22}$	$\mathbf{579.09_{\pm 10.42}}$	$600.44_{\pm 17.48}$
Average	120.60	421.204	423.56	419.47	442.27	**484.66**	535.74

Experiment: Ablation Study

- **Figure 2: Episodic accumulated returns.**



- **Table 3: Ablation: The impact of prompt decomposition, prompt distillation and test time adaptation.**

Decomposition	Distillation	TTA	Cheetah-vel	MW ML45	MW MT50
✗	✗	✗	-171.23	91.97	400.71
✗	✔	✔	-163.05	108.01	709.81
✔	✗	✔	-160.10	273.99	1137.39
✔	✔	✗	-167.80	149.21	824.07
✔	✔	✔	**-139.88**	**347.21**	**1559.94**

Experiment: Ablation Study

- **Table 4: Ablation: The impact of model size.**

Model size	Cheetah-vel	Ant-dir	MW ML45	MW MT50
(3,1,128)	-164.88	129.34	288.14	749.18
(12,12,768)	**-139.88**	121.84	**347.21**	**1559.94**
(24,16,768)	-210.35	**165.99**	292.48	1527.34

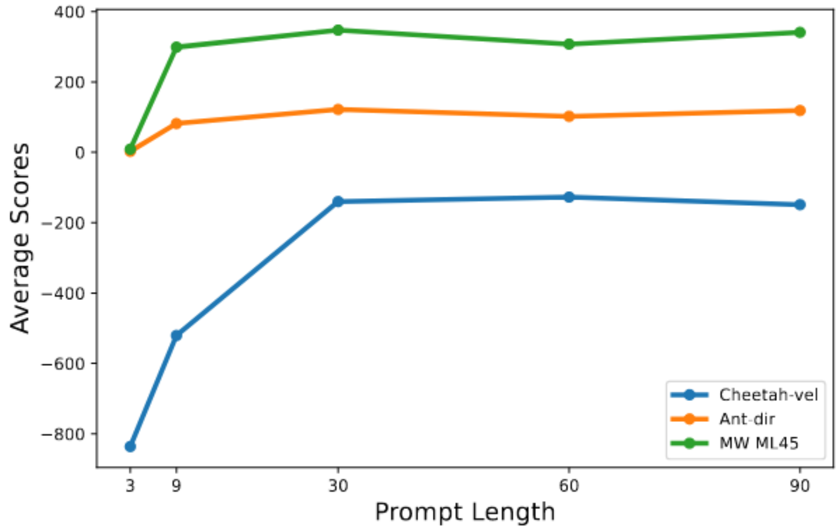
- **Table 5: Ablation: The impact of data quality.**

	Cheetah-vel	ML45
expert datasets	-30.10	586.84
medium datasets	-41.73	502.64
random datasets	-935.66	37.91
mixed datasets	-30.73	579.09

- **Table 6: Ablation: The impact of learning rate in prompt decomposition.**

	lr_{P_c}=1e-2	lr_{P_c}=1e-3	lr_{P_c}=1e-4
lr_{P_k}=1e-2	310.74	307.36	311.40
lr_{P_k}=1e-3	198.17	350.99	338.21
lr_{P_k}=1e-4	204.94	104.07	347.21

- **Figure 3: Ablation: The effect of prompt length on DPDT's zero-shot generalization ability.**



Thanks!

hlzheng@whu.edu.cn