Task-Aware Virtual Training: Enhancing Generalization in Meta-Reinforcement Learning for Out-of-Distribution Tasks

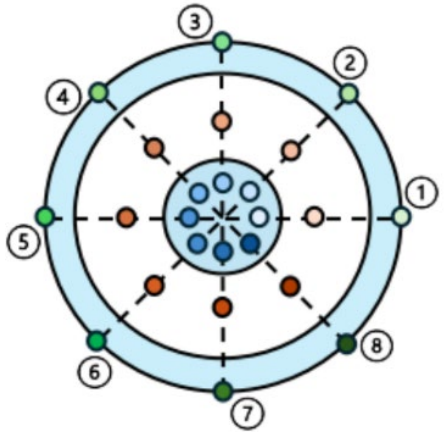
Jeongmo Kim, Yisak Park, Minung Kim, Seungyul Han*
UNIST, Artificial Intelligence Graduate School (AIGS)
{ jmkim22, isaac1018, minungkim, syhan }@unist.ac.kr

ICML
International Conference
On Machine Learning

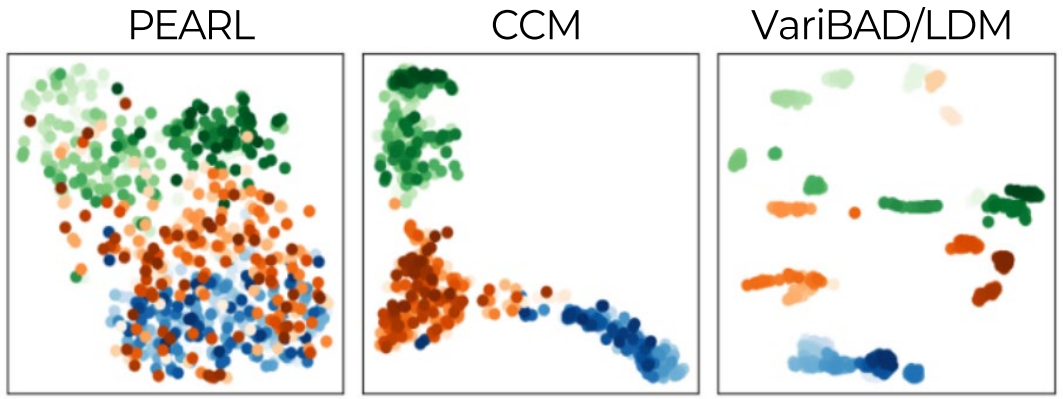
Motivation

❖ Insufficient task representation

➢ Ant-Goal-OOD



➢ Task Structure：
Goal Direction
Goal Distance



➢ Existing methods often fail to reflect the task structure!

Motivation：learning to well preserve the task structure even in OOD tasks

➢ Metric based representation learning

➢ Virtual task learning

Contribution1 - Metric Based Task Representation

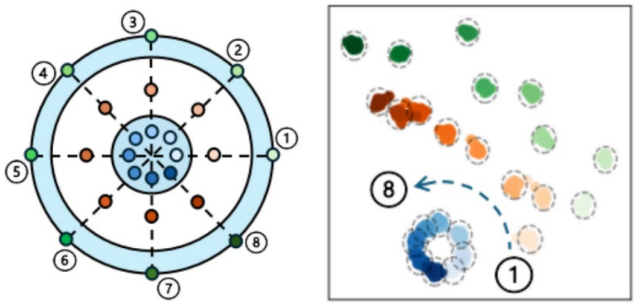
❖ Task metric design based on Bisimulation metric

➢ $$d(\mathcal{T}_i, \mathcal{T}_j) = \mathop{\mathbb{E}}_{(s,a)\sim D}\Big[|R^{\mathcal{T}_i}(s,a) - R^{\mathcal{T}_j}(s,a)| + \eta W_2(P^{\mathcal{T}_i}(\cdot|s,a), P^{\mathcal{T}_j}(\cdot|s,a))\Big] \quad (1)$$

❖ Metric based task representation learning $\mathcal{L}_{\text{bisim}}$

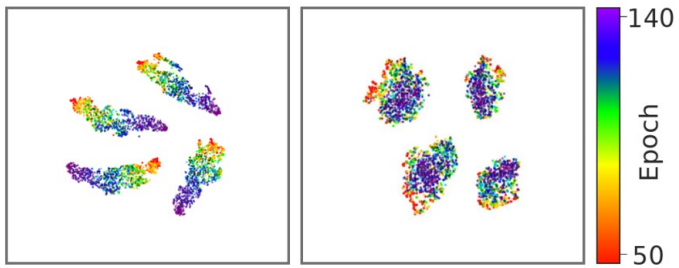
➢ $$\mathcal{L}_{\text{bisim}}(\psi,\phi) = \mathbb{E}_{\mathcal{T}_i,\mathcal{T}_j\sim p(\mathcal{T}_{\text{train}})}\Big[\underbrace{\Big(|\mathbf{z}^i_{\text{off}} - \mathbf{z}^j_{\text{off}}| - d(\mathcal{T}_i,\mathcal{T}_j;p_{\bar{\phi}})\Big)^2}_{\text{Bisimulation loss}} + \underbrace{\mathbb{E}_{(s,a,r,s')\sim D^{\mathcal{T}_i}_{\text{off}},(\hat{r},\hat{s}')\sim p_\phi(s,a,\mathbf{z}^i_{\text{off}})}\Big[(r-\hat{r})^2 + (s'-\hat{s}')^2\Big]}_{\text{Reconstruction loss}} + \underbrace{(\mathbf{z}^i_{\text{on}} - \bar{\mathbf{z}}^i_{\text{off}})^2}_{\text{on-off latent loss}} \Big] \quad (2)$$

➢ learned task latents well reflect task structure!



➢ train decoder to utilize task metric measuring

➢ train stable task latent $\mathbf{z}_{\text{on}}$



w/o on-off loss with on-off loss

Zhang et.al., Learning Invariant Representations for Reinforcement Learning without Reconstruction, ICLR 2021

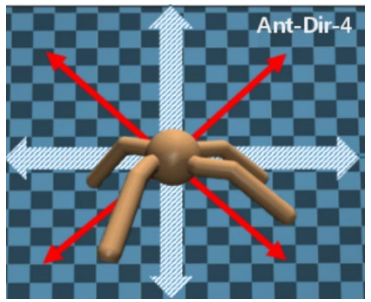
Contribution2 - Task Aware Sample Generation of Virtual Tasks

❖ Task aware virtual task sample generation and task preserving learning $\mathcal{L}_{\text{gen}}$

$$\mathcal{L}_{\text{gen}}(\psi, \phi) = \underbrace{\mathbb{E}_{\hat{\mathbf{c}}^\alpha \sim p_\phi}[-f_\zeta(\hat{\mathbf{c}}^\alpha, \bar{\mathbf{z}}^\alpha_{\text{off}})}_{\text{WGAN generator loss}} + \underbrace{\mathbb{E}_{\hat{\mathbf{z}}^\alpha \sim q_\psi(\cdot|\hat{\mathbf{c}}^\alpha)}[(\hat{\mathbf{z}}^\alpha - \bar{\mathbf{z}}^\alpha_{\text{off}})^2]]}_{\text{task preserving loss}} \quad (3)$$

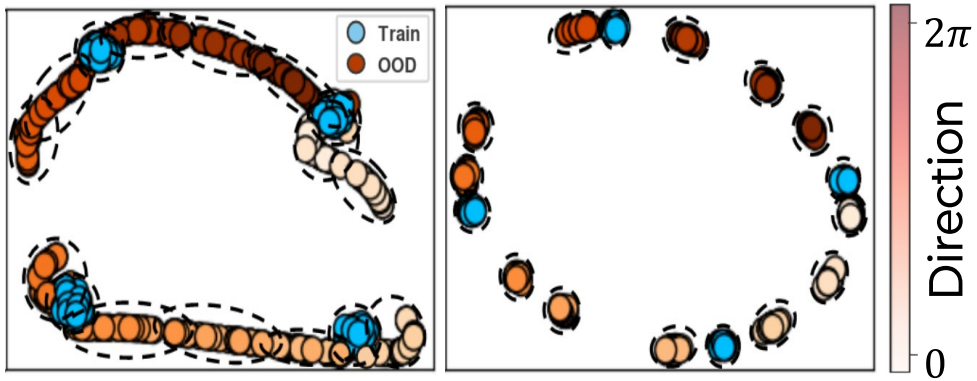
➢ enables to generate realistic task samples

➢ learn task latent to preserve task information of virtual task context



➢ Ant-Dir-4

Train task
Test task

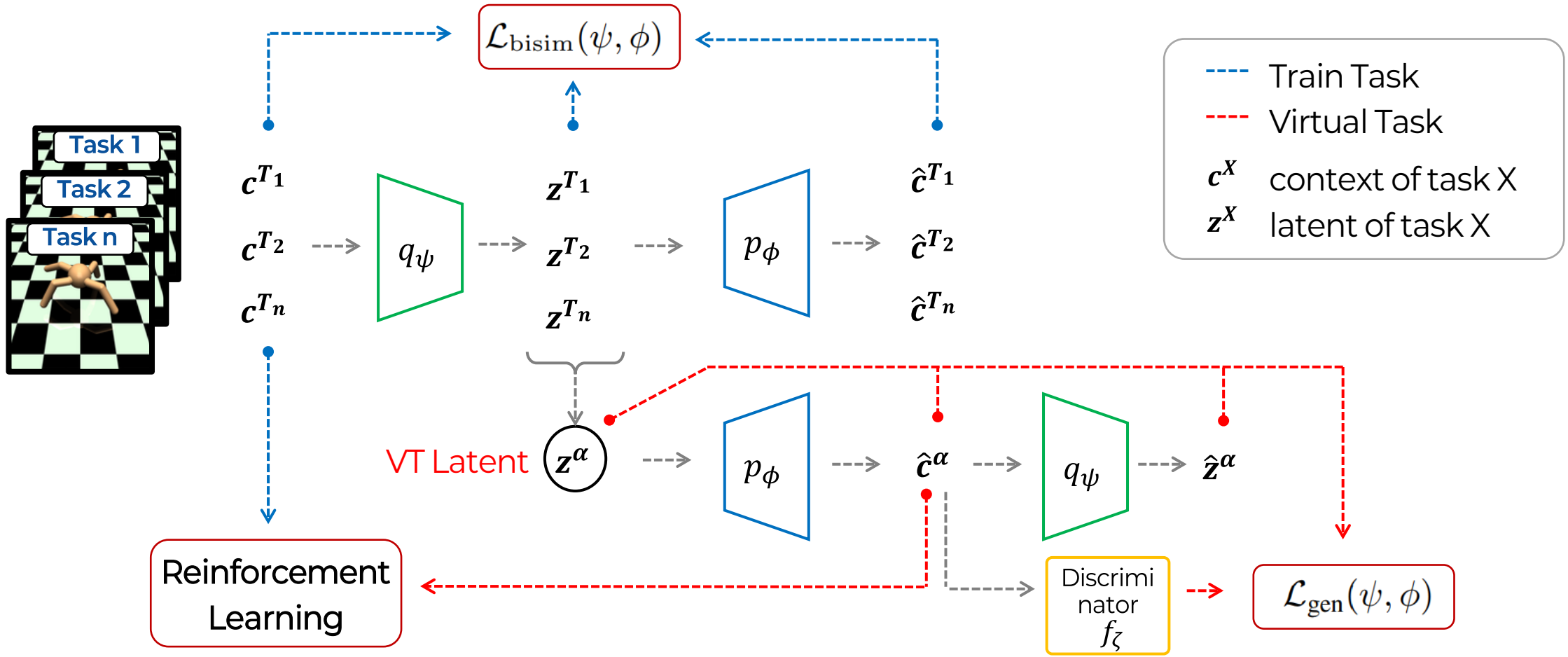


w/o $\mathcal{L}_{\text{gen}}$ with $\mathcal{L}_{\text{gen}}$

WGAN : Gulrajani et.al., Improved Training of Wasserstein GANs, NeurIPS 2017

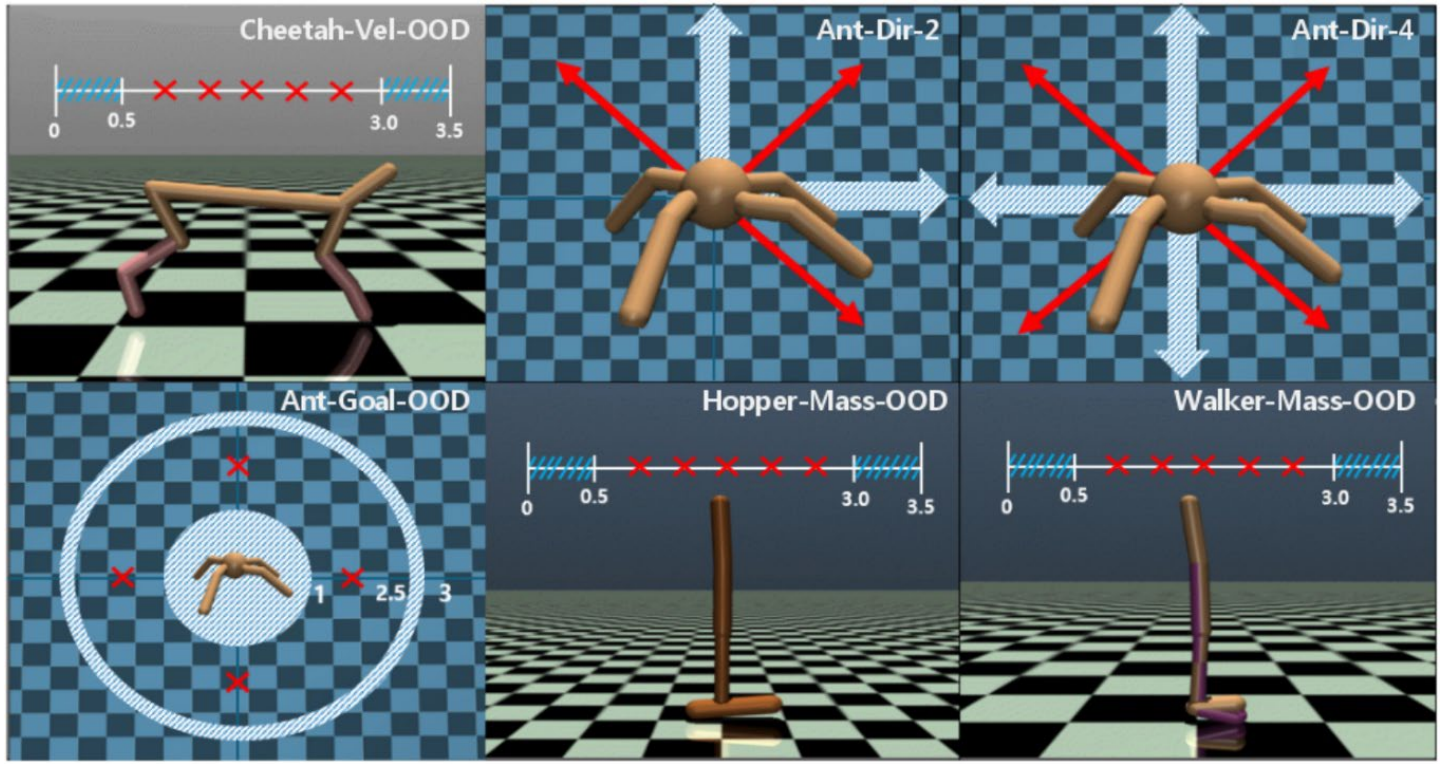
Method - Task Aware Virtual Training

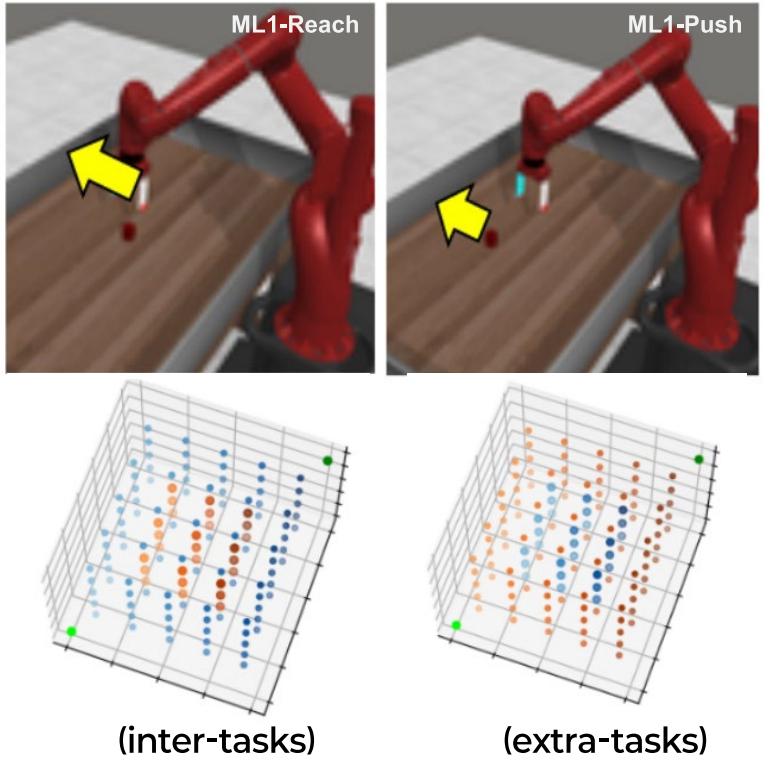
❖ Agent can prepare OOD tasks in advance through TAVT



Environments with OOD test tasks

❖ Environmental setup with Out-Of-Distribution test tasks





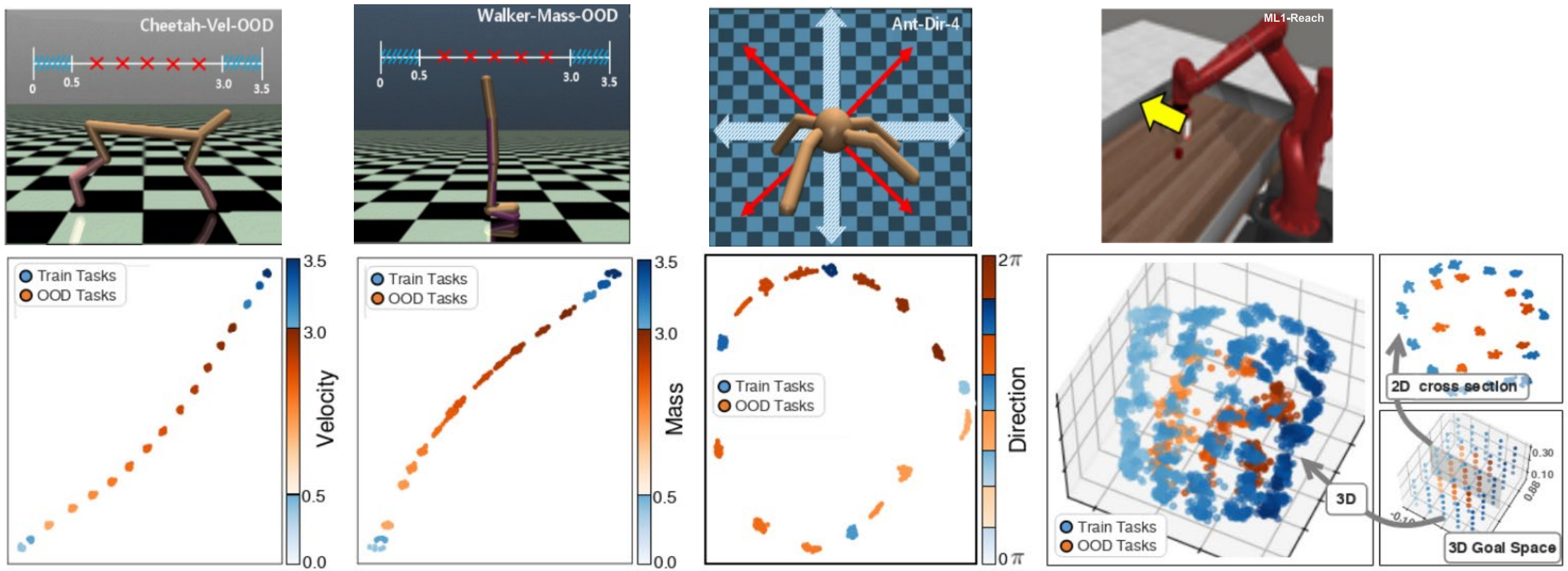
➢ MuJoCo Locomotion Task Environments

➢ MetaWorld ML1 Task Environments

Train task
Test task

Visualization of Task Representation

- TAVT well preserves all 1D, 2D and 3D task structure!



1-D (reward fn. varying)

1-D (transition fn. varying)

2-D

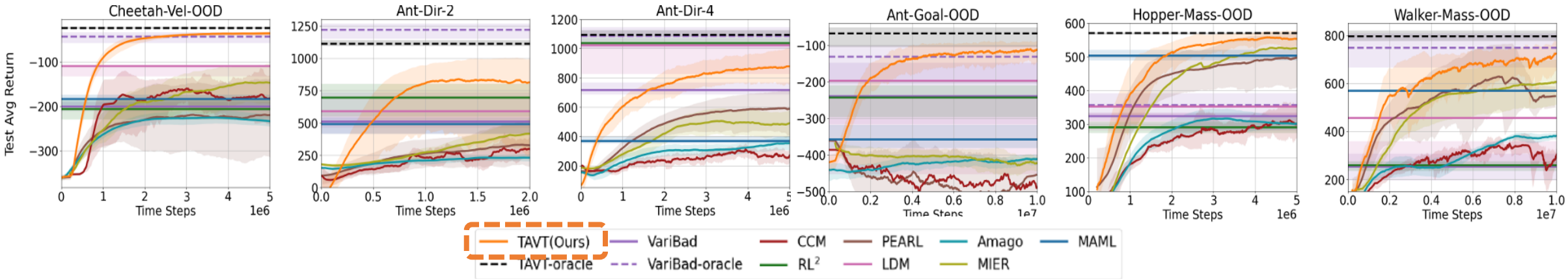
3-D

Train task
Test task

Performance on Out-Of-Distribution Tasks

❖ TAVT shows superior adaptation performance on various OOD test tasks!

➢ MuJoCo Locomotion Task Environments



➢ MetaWorld ML1 Task Environments

Table 3. Average success rate for MetaWorld ML1 environments.

	MAML	RL^2	VariBAD	LDM	PEARL	CCM	Amago	MIER	TAVT(Ours)
Reach	0.97±0.02	0.95±0.04	0.73±0.12	0.76±0.1	0.48±0.21	0.65±0.13	0.71±0.27	0.61±0.18	**0.98±0.02**
Reach-OOD-Inter	0.56±0.11	0.86±0.12	0.82±0.11	0.87±0.1	0.52±0.16	0.78±0.1	0.93±0.05	0.62±0.18	**0.96±0.03**
Reach-OOD-Extra	0.48±0.15	0.73±0.14	0.82±0.11	0.79±0.15	0.48±0.14	0.81±0.12	0.43±0.08	0.65±0.12	**0.99±0.01**
Push	0.94±0.03	**0.98±0.02**	0.88±0.09	0.83±0.11	0.61±0.11	0.18±0.08	0.87±0.11	0.59±0.13	**0.98±0.03**
Push-OOD-Inter	0.78±0.13	0.79±0.14	0.83±0.11	0.77±0.13	0.79±0.16	0.12±0.03	**0.98±0.02**	0.45±0.15	**0.98±0.02**
Push-OOD-Extra	0.55±0.13	0.38±0.12	0.65±0.09	0.72±0.11	0.55±0.18	0.15±0.04	0.83±0.11	0.61±0.15	**0.92±0.08**

Thank you!

See you soon on the Conference!