

A TRAINING APPROACH AND PSEUDOCODE

We utilize Proximal Policy Optimization (PPO) (Schulman et al., 2017) to train our policy network. PPO is an on-policy, actor-critic deep RL algorithm. The optimization objective for the policy is as follows:

$$\mathcal{L}_{\text{policy}}(\theta) = \mathbb{E} \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \right] \quad (9)$$

Here, \hat{A}_t denotes the estimation of the advantage function, and $r_t(\theta)$ represents the probability ratio, defined as $r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, where π_{θ} represents the new policy and $\pi_{\theta_{\text{old}}}$ represents the old policy.

In optimizing the conditional policy module, we utilize the $\mathcal{L}_{\text{policy}}$ loss function. Notably, similar to the approach utilized in VariBAD (Zintgraf et al., 2019), the optimization of the task inference module does not rely on the $\mathcal{L}_{\text{policy}}$ loss function. Instead, we adopt a composite loss function that combines reconstruction and contrastive learning objectives. The specific pseudo-code is shown in Algorithm 1.

Algorithm 1 CFOHLR

Require: Encoder q_{ϕ} and Decoder p_{θ} of VAE; Coarse policy π_{θ} ; Fine policy π_{ω} ; Weight generator W_{α} Skill-specific expert models $\text{MoE}_{\theta_i}^{i=0,\dots,K}$; Hypernetworks H_{ϕ} ; VAE buffer \mathcal{D}_{VAE} ; Policy buffer $\mathcal{D}_{\text{Policy}}$; The number of meta-episodes n_{meta} ; The number of rollout episodes per meta-episode n_{roll} ; Language instruction \mathcal{U} .

while $i=0,\dots,N_{\text{update}}$ **do**

 Sample K training tasks $M_i^{i=0,\dots,K} \sim M_{\text{train}}$

for timestep $t = 0,\dots,n_{\text{roll}} * H - 1$ **do**

if $t \bmod H = 0$ **then**

 Reset rollout episode for each task, obtain $S_t = \{s_{t,1}, s_{t,2}, \dots, s_{t,n}\}$

end if

for $j=0,\dots,K$ **do**

 Obtain weights $\alpha_{1,j}, \dots, \alpha_{K,j} = W_{\alpha}(\mathcal{U}_j)$ for each skill-specific expert module.

 Obtain the output of the coarse policy π_{θ} , denoted as $O_{\text{MOE}} = \sum_{i=0}^K \alpha_i \cdot \text{MoE}_{\theta_i}$.

 Leverage H_{ϕ} to generate the network parameters of the fine policy, $\pi_{\omega} = H_{\phi}(z_t^j)$.

 Obtain the action $a_{t,j} = \pi_{\omega}(O_{\text{MOE}})$.

end for

 Finally, obtain actions for each task $A_t = \{a_{t,1}, a_{t,2}, \dots, a_{t,n}\}$.

 Take an environment step, obtaining $S_{t+1} = \{s_{t+1,1}, s_{t+1,2}, \dots, s_{t+1,n}\}$ and $R_t = \{r_{t+1,1}, r_{t+1,2}, \dots, r_{t+1,n}\}$.

 Add the transition $(S_t, A_t, R_{t+1}, S_{t+1})$ to \mathcal{D}_{VAE} and $\mathcal{D}_{\text{Policy}}$.

 Update task representations $Z_{t+1} = \{z_{t+1,n} = q_{\phi}(\tau_{t+1,n})\}_{i=0,\dots,n}$.

end for

 Update VAE by minimizing $\mathcal{L} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{contra}}$

 Update policy θ, ω and weight generator α by minimizing $L_{\text{actor}} + L_{\text{critic}}$.

end while

B LIMITATIONS AND FUTURE WORK

Despite the significant progress, our method has limitations that were not addressed in this study. Notably, it is not directly applicable to the cross-entity adaptation problem, which involves generalizing a policy from one robotic entity to another. This limitation affects the overall generalizability of the policy. Future research will focus on tackling the challenge of cross-entity adaptation in a zero-shot manner, thereby enhancing the policy generalization.

C IMPLEMENTATION DETAILS

C.1 REFERENCE IMPLEMENTATIONS

SDVT, LDM, and VariBAD We adapt the SDVT (Lee et al., 2023), LDM (Lee & Chung, 2021), and VariBAD (Zintgraf et al., 2019) algorithms to the Meta-World benchmark. These algorithms are all based on the VariBAD method, which itself is grounded in the Bayesian Adaptive MDP (BAMDP) framework. VariBAD employs a VAE architecture consisting of a recurrent encoder and a dynamics decoder to obtain task representations. LDM introduces a virtual training procedure to VariBAD to address out-of-distribution challenges. Building on LDM, SDVT uses a Gaussian mixture distribution to model the latent space of the VAE. Notably, the virtual training steps of the LDM and SDVT methods are included in the total count of training steps, as these virtual processes necessitate agent interaction with the environment to obtain real states for generating imagined samples. We used open-source code to reproduce the results of the SDVT, LDM, and VariBAD methods, respectively, available at <https://github.com/suyoung-lee/SDVT>, <https://github.com/suyoung-lee/LDM>, and <https://github.com/lmzintgraf/varibad>.

Million Million (Bing et al., 2023) introduces a meta-RL paradigm comprising an instruction phase and a trial phase, integrating transformers with language instruction to improve task adaptation capabilities. We used open-source code to reproduce the results of the Million methods, respectively, available at <https://github.com/yaopt3/MILLION>.

C.2 HYPERPARAMETERS

C.2.1 SDVT

We used the default hyperparameters from the paper, which are shown in Table 3.

C.2.2 LDM AND VARIBAD

We used the default hyperparameters from the paper, which are shown in Table 4.

C.2.3 MILLION

We used the default hyperparameters from the paper, which are shown in Table 5.

C.2.4 OURS

C.3 NETWORK ARCHITECTURE

Our method utilizes a context-based architecture, comprising a task inference module and a conditional policy module. For the task inference module, similar to SDVT, we also employ a Gaussian mixture VAE to model the latent space. This module consists of an RNN-based encoder and a prediction decoder. Before being input into the encoder or decoder, all state, action, and reward inputs pass through embedding networks. Regarding the conditional policy module, it includes language-selected, skill-specific expert modules and a hypernetwork-based, task-aware policy. Similarly, before being input into the conditional policy module, all state, action, and reward inputs pass through embedding networks.

C.4 TASK DESCRIPTIONS

In Table 11, we provide the language instructions for each of the 50 Meta-World tasks.

D DETAILED EXPERIMENTAL RESULTS

We adhere to the success criterion established by Meta-World. A timestep is considered successful when the distance between the task-relevant object and the target falls within an acceptable range. Furthermore, an entire rollout episode is deemed successful if the agent achieves success at any timestep during the episode.

Table 3: Hyperparameters used for Garage experiments with SDVT

Description	ML10	ML45
Meta-Task Hyperparameters		
Meta-batch size	10	10
Tasks sampled per epoch	10	10
General Hyperparameters		
Batch size	1,000	1,000
Path length per roll-out	1,000	1,000
Discount factor	0.99	0.99
Algorithm-Specific Hyperparameters		
Policy hidden sizes	(256, 256)	(256, 256)
Activation function	tanh	tanh
Policy learning rate	7×10^{-4}	7×10^{-4}
PPO epochs num	5	5
VAE learning rate	1×10^{-3}	1×10^{-3}
Latent dimension	5	5
PPO num minibatches	10	10
PPO clip param	0.1	0.1
Policy num steps	5	5
Size of VAE buffer	1,000	1,000
KL weight	0.1	0.1
VAE mixture num	10	10
Gaussian loss coefficient	1.0	1.0
Action embedding size	16	16
State embedding size	32	32
Reward embedding size	16	16
Virtual ratio increment	0.05	0.05
Number of virtual skills	3	3
RL loss through encoder	False	False
VAE loss coefficient	1.0	1.0

D.1 PERFORMANCE ON INDIVIDUAL TASKS

D.1.1 ML-10

D.1.2 ML-45

D.2 LEARNING CURVES

In Figure 5, we present the mean and standard deviation of returns and success rates across five random seeds.

E ADDITIONAL RESULTS

E.1 VISULIZATIONS

To demonstrate the quality of the learned task representations, we employed t-SNE Van der Maaten & Hinton (2008) to map the task representation vectors into a 2D space, enabling the visualization of these representations. For each testing task, 150 transitions from the meta-testing phase were sampled to visualize the task representations. As depicted in Figure 6, our method effectively distinguishes task representations from different categories, with additional separation observed among tasks within the same category.

Table 4: The hyperparameters used in experiments with LDM and VariBAD are consistent across both models in the general and policy categories of SDVT, as outlined in Table 3. The only difference lies in the modeling of the latent space: SDVT utilizes a Gaussian mixture model, while both LDM and VariBAD employ a Gaussian model.

Description	ML10	ML45
Meta-Task Hyperparameters		
Meta-batch size	10	10
Tasks sampled per epoch	10	10
General Hyperparameters		
Batch size	1,000	1,000
Path length per roll-out	1,000	1,000
Discount factor	0.99	0.99
Algorithm-Specific Hyperparameters		
VAE learning rate	1×10^{-3}	1×10^{-3}
Latent dimension	5	5
Size of VAE buffer	1,000	1,000
KL weight	0.1	0.1
Gaussian loss coefficient	1.0	1.0
VAE loss coefficient	1.0	1.0

Table 5: Hyperparameters used in experiments with Million.

Description	ML10	ML45
Meta-Task Hyperparameters		
Meta-batch size	10	10
Tasks sampled per epoch	10	10
General Hyperparameters		
Batch Timesteps	1,000	1,000
Action repeat	1,000	1,000
Demonstration action repeat	1,000	1,000
Max trials per episode	750	750
Discount factor	0.99	0.99
Algorithm-Specific Hyperparameters		
Learning rate	$1e - 4$	$1e - 4$
GAE lambda	0.97	0.97
Epsilon eta	1×10^{-2}	1×10^{-2}
Epsilon alpha	1×10^{-2}	1×10^{-2}
Epsilon alpha mu	0.0075	0.0075
Epsilon alpha sigma	$1e - 5$	$1e - 5$

Table 6: Hyperparameters used in experiments with Ours.

Description	ML10	ML45
Meta-Task Hyperparameters		
Meta-batch size	10	10
Tasks sampled per epoch	10	10
General Hyperparameters		
Batch size	1,000	1,000
Path length per roll-out	1000	
Discount factor	0.99	
Algorithm-Specific Hyperparameters		
Policy hidden sizes	(256, 256)	(256, 256)
Activation function	tanh	tanh
Policy learning rate	7×10^{-4}	7×10^{-4}
PPO epochs num	5	5
VAE learning rate	1×10^{-3}	1×10^{-3}
Latent dimension	5	5
PPO num minibatches	10	10
PPO clip param	0.1	0.1
Policy num steps	5	5
RL loss through encoder	False	False
Action embedding size	16	16
State embedding size	32	32
Reward embedding size	16	16
Size of VAE buffer	1,000	1,000
KL weight	0.1	0.1
VAE mixture num	10	10
Gaussian loss coefficient	1.0	1.0
VAE loss coefficient	1.0	1.0
Decode reward	True	True
Decode state	True	True
Weight of holistic	0.01	0.01
Weight of local contrastive	0.01	0.01

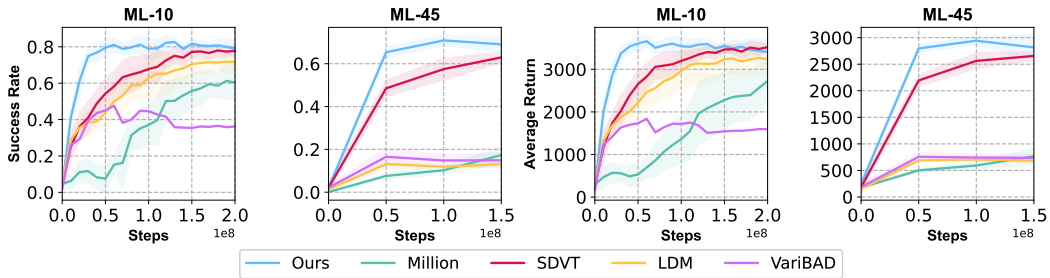


Figure 5: **Learning Curves on ML-10 and ML-45.** The maximum success rates and corresponding returns of our methods, along with baseline comparisons, are presented. The plot shows the mean and standard deviation of returns across five random seeds.

Table 7: ML-10 of Meta-World success rate (%). We present the final success rates of our method and baseline approaches on both the training and test tasks of the Meta-World ML-10 benchmark. All results are reported as the mean success rate $\pm 95\%$ confidence interval of five seeds.

Index. Task	Ours	w/o C2F	w/o HLC	SDVT	Million	LDM	VariBAD
1. Reach	85.2\pm3.6	53.2 \pm 4.6	44.0 \pm 7.3	53.6 \pm 14.4	10.4 \pm 11.3	50.4 \pm 7.9	78.0 \pm 5.7
2. Push	86.0\pm3.0	70.4 \pm 9.8	74.8 \pm 3.0	74.0 \pm 3.8	44.5 \pm 17.8	31.2 \pm 11.2	2.4 \pm 2.0
3. Pick-place	85.2\pm4.5	66.0 \pm 6.9	66.0 \pm 2.9	53.2 \pm 6.0	48.1 \pm 29.8	37.2 \pm 20.0	0.8 \pm 0.7
4. Door-open	81.2 \pm 1.9	99.6 \pm 0.6	97.2 \pm 3.9	100.0\pm0.0	81.1 \pm 25.1	99.6 \pm 0.6	74.8 \pm 25.4
5. Drawer-close	86.8 \pm 1.4	100.0\pm0.0	100.0\pm0.0	100.0\pm0.0	56.1 \pm 32.1	100.0\pm0.0	100.0\pm0.0
6. Button-press	84.4 \pm 3.1	100.0\pm0.0	100.0\pm0.0	99.6 \pm 0.6	80.0 \pm 27.7	98.4 \pm 1.0	88.4 \pm 4.1
7. Peg-insert-side	88.0\pm2.5	48.8 \pm 18.6	62.0 \pm 13.1	52.0 \pm 5.0	21.5 \pm 18.3	26.4 \pm 15.3	0.0 \pm 0.0
8. Window-open	86.4 \pm 4.4	99.6 \pm 0.6	100.0\pm0.0	100.0\pm0.0	80.0 \pm 27.7	99.6 \pm 0.6	96.8 \pm 1.1
9. Sweep	89.6 \pm 2.7	93.2\pm2.1	93.2\pm3.5	89.2 \pm 3.8	77.6 \pm 24.0	92.4 \pm 3.3	0.0 \pm 0.0
10. Basketball	82.8 \pm 4.5	93.6\pm4.0	92.0 \pm 5.6	72.8 \pm 15.5	36.1 \pm 31.0	89.2 \pm 4.0	0.0 \pm 0.0
Train mean	85.6\pm3.9	82.4 \pm 7.7	82.9 \pm 4.8	79.4 \pm 7.1	53.5 \pm 40.0	72.4 \pm 11.4	44.1 \pm 7.9
11. Drawer-open	86.4 \pm 3.4	78.0 \pm 24.1	80.0 \pm 12.6	72.8 \pm 28.5	0.0 \pm 0.0	92.8\pm8.1	51.6 \pm 21.1
12. Door-close	87.2 \pm 3.8	74.4 \pm 25.7	81.6 \pm 15.7	76.4 \pm 19.7	97.5\pm2.8	26.0 \pm 37.7	71.6 \pm 22.4
13. Shelf-place	82.4\pm5.5	1.6 \pm 3.1	0.4 \pm 0.8	0.0 \pm 0.0	0.3 \pm 0.5	0.4 \pm 0.8	0.0 \pm 0.0
14. Sweep-into	90.0\pm5.1	67.6 \pm 34.0	82.8 \pm 8.8	64.8 \pm 12.2	17.5 \pm 5.0	57.6 \pm 31.9	4.8 \pm 3.2
15. Lever-pull	82.4\pm3.4	5.2 \pm 7.4	0.4 \pm 0.8	3.2 \pm 3.2	13.7 \pm 26.9	0.4 \pm 0.8	0.4 \pm 0.8
Test mean	85.7\pm4.9	45.4 \pm 13.7	49.0 \pm 5.5	43.4 \pm 9.4	25.8 \pm 9.1	35.4 \pm 17.1	25.7 \pm 7.5

Table 8: ML-10 of Meta-World returns. We present the performance metrics of our method and baseline approaches on both the training and test tasks of the Meta-World ML-10 benchmark. All results are reported as the mean return $\pm 95\%$ confidence interval of five seeds.

Index. Task	Ours	w/o C2F	w/o HLC	SDVT	Million	LDM	VariBAD
1. Reach	3704 \pm 149	3778\pm128	3520 \pm 141	3763 \pm 296	2324 \pm 447	3668 \pm 285	4054 \pm 138
2. Push	3769\pm135	3338 \pm 342	4094 \pm 90	3675 \pm 272	2225 \pm 750	1795 \pm 812	63 \pm 28
3. Pick-place	3742\pm127	2089 \pm 254	2420 \pm 116	1712 \pm 125	1678 \pm 803	1258 \pm 709	7 \pm 1
4. Door-open	3740 \pm 91	4503 \pm 51	4313 \pm 78	4439 \pm 26	2790 \pm 980	4442\pm47	2978 \pm 470
5. Drawer-close	3708 \pm 75	4857\pm7	4811 \pm 30	4852 \pm 10	2505 \pm 1302	4809 \pm 67	4637 \pm 77
6. Button-press	3622 \pm 144	3489 \pm 93	3250 \pm 108	3372\pm60	1337 \pm 734	3226 \pm 60	2028 \pm 155
7. Peg-insert-side	3703\pm113	2359 \pm 678	2827 \pm 421	2443 \pm 266	1179 \pm 697	1364 \pm 773	9 \pm 1
8. Window-open	3787 \pm 157	4479\pm49	4398 \pm 49	4476 \pm 51	2331 \pm 942	4384 \pm 61	3692 \pm 202
9. Sweep	3786 \pm 147	4093\pm85	3963 \pm 100	3801 \pm 208	2849 \pm 932	3997 \pm 189	92 \pm 28
10. Basketball	3705\pm185	3532 \pm 196	3576 \pm 202	2937 \pm 618	1624 \pm 774	3433 \pm 196	9 \pm 2
Train mean	3727\pm221	3652 \pm 289	3717 \pm 112	3547 \pm 203	2084 \pm 1365	3238 \pm 613	1757 \pm 178
11. Drawer-open	3796\pm112	2477 \pm 393	2477 \pm 190	2660 \pm 396	1876 \pm 384	2697 \pm 475	2036 \pm 329
12. Door-close	3740\pm119	2489 \pm 566	2887 \pm 769	3087 \pm 944	3302 \pm 415	1272 \pm 1538	2113 \pm 558
13. Shelf-place	3746\pm157	492 \pm 246	607 \pm 115	341 \pm 99	141 \pm 204	309 \pm 272	0 \pm 0
14. Sweep-into	3751\pm128	1619 \pm 786	1705 \pm 434	1444 \pm 564	716 \pm 264	1200 \pm 793	172 \pm 96
15. Lever-pull	3634\pm197	305 \pm 44	251 \pm 29	285 \pm 60	208 \pm 44	278 \pm 59	324 \pm 41
Test mean	3734\pm165	1476 \pm 224	1585 \pm 290	1563 \pm 418	1249 \pm 205	1151 \pm 692	929 \pm 208

Table 9: ML-45 of Meta-World success rate (%). We present the final success rates of our method and baseline approaches on both the training and test tasks of the Meta-World ML-45 benchmark. All results are reported as the mean success rate $\pm 95\%$ confidence interval of five seeds.

Index. Task	Ours	w/o C2F	w/o HLC	SDVT	Million	LDM	VariBAD
1. Assembly	68.0\pm2.0	0.5 \pm 0.3	0.5 \pm 0.3	0.5 \pm 0.3	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
2. Basketball	67.5\pm1.3	22.0 \pm 5.7	21.5 \pm 1.7	38.0 \pm 5.2	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
3. Button-press-topdown	65.0 \pm 3.2	99.0\pm0.6	98.5 \pm 0.3	98.0 \pm 0.7	0.0 \pm 0.0	9.5 \pm 2.8	27.5 \pm 6.7
4. Button-press-topdown-wall	72.0 \pm 2.3	96.0 \pm 2.3	98.5 \pm 0.6	99.0\pm0.3	0.0 \pm 0.0	9.5 \pm 2.7	17.0 \pm 4.6
5. Button-press-wall	75.5 \pm 3.0	94.0 \pm 2.1	95.0 \pm 1.8	100.0\pm0.0	43.2 \pm 10.8	30.0 \pm 3.9	27.0 \pm 6.1
6. Button-press-wall	71.0 \pm 2.4	88.0\pm3.1	76.0 \pm 4.8	80.5 \pm 4.2	49.2 \pm 6.0	41.0 \pm 9.7	32.0 \pm 4.8
7. Coffee-button	65.0 \pm 2.8	100.0\pm0.0	100.0\pm0.0	98.5 \pm 0.9	42.3 \pm 14.7	55.5 \pm 11.0	33.5 \pm 8.7
8. Coffee-pull	73.5\pm2.9	62.5 \pm 1.3	70.0 \pm 3.6	37.0 \pm 5.8	0.2 \pm 0.1	1.0 \pm 0.3	0.0 \pm 0.0
9. Coffee-push	76.5\pm1.5	57.5 \pm 10.7	69.5 \pm 5.5	37.5 \pm 4.9	30.8 \pm 9.7	11.5 \pm 3.3	7.5 \pm 2.1
10. Dial-turn	73.5 \pm 2.6	73.0 \pm 6.1	75.0 \pm 4.9	77.5\pm2.6	62.2 \pm 2.4	8.5 \pm 1.7	26.0 \pm 6.4
11. Disassemble	74.0 \pm 3.2	69.5 \pm 13.6	79.0\pm5.3	78.0 \pm 4.3	0.0 \pm 0.0	0.5 \pm 0.3	4.5 \pm 2.6
12. Door-close	72.5 \pm 1.3	100.0\pm0.0	99.5 \pm 0.3	100.0 \pm 0.0	51.0 \pm 11.9	98.5 \pm 0.9	65.0 \pm 13.8
13. Door-open	74.5 \pm 2.8	92.0 \pm 3.3	94.0 \pm 2.7	95.5\pm0.9	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
14. Drawer-close	70.5 \pm 1.7	96.0 \pm 1.4	100.0\pm0.0	99.5 \pm 0.3	100.0\pm0.0	99.0 \pm 0.6	100.0\pm0.0
15. Drawer-open	68.0 \pm 2.4	99.5\pm0.3	95.0 \pm 1.5	98.5 \pm 0.6	0.0 \pm 0.0	15.0 \pm 6.1	10.0 \pm 2.4
16. Faucet-open	71.0 \pm 2.7	98.0 \pm 0.5	96.0 \pm 2.0	98.5\pm0.9	11.2 \pm 3.8	30.5 \pm 1.5	54.0 \pm 9.7
17. Faucet-close	77.0 \pm 1.0	98.5 \pm 0.9	87.5 \pm 3.8	99.5\pm0.3	42.0 \pm 13.5	22.0 \pm 4.4	28.5 \pm 7.1
18. Hammer	75.5\pm1.5	8.5 \pm 5.0	33.0 \pm 11.1	0.0 \pm 0.0	12.2 \pm 7.1	1.5 \pm 0.9	2.0 \pm 0.8
19. Handle-press-side	69.0 \pm 1.4	100.0\pm0.0	99.0 \pm 0.6	100.0\pm0.0	14.2 \pm 4.9	6.0 \pm 2.2	38.5 \pm 8.5
20. Handle-press	70.0 \pm 1.3	100.0\pm0.0	99.5 \pm 0.3	100.0\pm0.0	65.5 \pm 3.2	45.5 \pm 2.9	57.0 \pm 2.5
21. Handle-pull-side	72.5 \pm 3.0	96.0 \pm 2.0	98.0\pm1.2	89.5 \pm 3.6	17.7 \pm 4.6	0.0 \pm 0.0	1.5 \pm 0.6
22. Handle-pull	70.5 \pm 1.9	76.5 \pm 13.3	99.5\pm0.3	63.5 \pm 12.2	27.0 \pm 10.1	1.5 \pm 0.3	1.0 \pm 0.6
23. Lever-pull	76.5\pm1.2	55.5 \pm 9.6	9.5 \pm 5.6	52.0 \pm 10.4	0.0 \pm 0.0	0.0 \pm 0.0	1.5 \pm 0.3
24. Peg-insert-side	72.0\pm1.7	9.0 \pm 1.1	33.5 \pm 2.5	3.5 \pm 0.6	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
25. Pick-place-wall	72.0\pm1.1	61.5 \pm 4.2	59.5 \pm 6.4	46.0 \pm 3.5	13.5 \pm 5.8	0.0 \pm 0.0	0.5 \pm 0.3
26. Pick-out-of-hole	67.0\pm3.6	39.0 \pm 9.9	52.0 \pm 5.0	53.5 \pm 5.7	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
27. Push	79.5\pm3.0	38.0 \pm 3.0	48.0 \pm 2.2	27.5 \pm 3.2	7.7 \pm 2.9	42.0 \pm 3.8	46.0 \pm 5.8
28. Push-back	66.5 \pm 1.9	69.0 \pm 5.7	79.5\pm3.1	64.0 \pm 4.2	0.0 \pm 0.0	0.0 \pm 0.0	1.0 \pm 0.6
29. Push	74.5\pm2.6	43.0 \pm 4.2	68.0 \pm 4.9	38.5 \pm 7.5	8.3 \pm 2.4	4.5 \pm 1.0	3.0 \pm 1.0
30. Pick-place	67.5\pm1.2	50.5 \pm 2.4	58.5 \pm 2.8	47.5 \pm 3.6	10.3 \pm 4.3	0.0 \pm 0.0	1.0 \pm 0.3
31. Plate-slide-side	69.0\pm2.8	47.5 \pm 4.8	48.0 \pm 3.7	67.0 \pm 3.8	36.7 \pm 7.0	0.0 \pm 0.0	0.5 \pm 0.3
32. Plate-slide-side	71.0 \pm 3.2	95.0\pm1.0	91.5 \pm 2.2	92.5 \pm 2.3	0.0 \pm 0.0	0.5 \pm 0.3	13.5 \pm 7.9
33. Plate-slide back	67.5 \pm 1.5	96.5 \pm 0.6	98.5\pm0.9	91.5 \pm 1.5	0.0 \pm 0.0	1.5 \pm 0.6	4.5 \pm 1.5
34. Plate-slide-back-side	75.5 \pm 2.4	80.5 \pm 5.1	89.0\pm2.5	83.5 \pm 2.7	0.0 \pm 0.0	0.5 \pm 0.3	6.0 \pm 2.0
35. Peg-unplug-side	69.5 \pm 1.7	66.5 \pm 3.7	76.0\pm4.7	53.5 \pm 3.8	5.3 \pm 1.3	6.5 \pm 2.8	4.5 \pm 1.5
36. Soccer	73.0\pm1.7	21.5 \pm 4.8	18.0 \pm 1.2	26.0 \pm 2.4	10.5 \pm 2.1	9.0 \pm 2.2	8.0 \pm 1.3
37. Stick-push	72.0\pm1.4	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
38. Stick-pull	73.5\pm1.9	0.0 \pm 0.0	0.5 \pm 0.3	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
39. Push-wall	72.0\pm2.4	54.5 \pm 6.4	88.0 \pm 1.1	54.0 \pm 10.0	7.8 \pm 3.1	0.5 \pm 0.3	1.0 \pm 0.6
40. Reach-wall	75.0\pm0.6	40.5 \pm 6.9	60.0 \pm 4.1	26.5 \pm 6.1	13.3 \pm 5.3	49.5 \pm 5.6	75.0\pm5.2
41. Shelf-place	74.5\pm2.5	6.0 \pm 2.8	1.0 \pm 0.6	1.0 \pm 0.6	0.8 \pm 0.5	0.0 \pm 0.0	0.0 \pm 0.0
42. Sweep-into	66.0 \pm 1.4	94.5 \pm 1.2	95.5\pm1.3	81.5 \pm 8.1	25.0 \pm 4.7	9.0 \pm 1.9	11.0 \pm 2.6
43. Sweep	73.0 \pm 1.1	74.5 \pm 2.9	83.0\pm3.6	36.5 \pm 12.3	30.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
44. Window-open	64.5 \pm 0.7	99.0\pm0.3	98.5 \pm 0.9	100.0 \pm 0.0	0.60 \pm 7.9	13.5 \pm 3.6	33.5 \pm 6.2
45. Window-close	67.0 \pm 2.7	100.0\pm0.0	99.5 \pm 0.3	99.5 \pm 0.3	11.8 \pm 4.0	17.0 \pm 3.2	29.5 \pm 8.0
Train mean	71.4\pm4.3	66.0 \pm 5.8	69.8 \pm 4.1	63.0 \pm 5.5	17.3 \pm 8.9	14.2 \pm 2.6	17.2 \pm 4.2
46. Bin-picking	76.0\pm3.2	1.0 \pm 1.0	1.5 \pm 0.9	3.5 \pm 2.6	0.2 \pm 0.3	0.0 \pm 0.0	0.0 \pm 0.0
47. Box-close	70.0\pm4.3	1.0 \pm 1.0	6.5 \pm 3.9	0.5 \pm 0.9	1.7 \pm 2.9	0.5 \pm 0.9	0.5 \pm 0.9
48. Hand-insert	69.5\pm2.6	0.5 \pm 0.9	2.5 \pm 2.6	3.5 \pm 3.6	7.5 \pm 7.7	3.0 \pm 3.4	3.0 \pm 2.3
49. Door-lock	73.0 \pm 10.0	82.0\pm3.8	70.0 \pm 17.2	61.5 \pm 13.4	34.3 \pm 4.8	41.5 \pm 10.1	134.5 \pm 12.6
50. Door-unlock	68.5 \pm 3.6	81.0 \pm 7.1	84.5\pm9.7	66.0 \pm 15.5	14.2 \pm 24.8	21.0 \pm 10.0	37.0 \pm 15.0
Test mean	71.4\pm3.5	33.1 \pm 3.0	33.0 \pm 6.3	27.0 \pm 8.9	11.6 \pm 7.1	13.2 \pm 6.0	15.0 \pm 6.5

Table 10: ML-45 of Meta-World returns. We present the final returns of our method and baseline approaches on both the training and test tasks of the Meta-World ML-45 benchmark. All results are reported as the mean return $\pm 95\%$ confidence interval of five seeds.

Index. Task	Ours	w/o C2F	w/o HLC	SDVT	Million	LDM	VariBAD
1. Assembly	2898\pm97	2847 \pm 27	2529 \pm 88	2590 \pm 61	329 \pm 53	154 \pm 27	101 \pm 13
2. Basketball	2885\pm101	1417 \pm 116	1444 \pm 107	1569 \pm 169	281 \pm 86	5 \pm 0	11 \pm 2
3. Button-press-topdown	2693 \pm 116	3582 \pm 100	3712\pm27	3586 \pm 50	884 \pm 128	988 \pm 104	1182 \pm 84
4. Button-press-topdown-wall	2950 \pm 69	3541 \pm 120	3686\pm52	3594 \pm 62	868 \pm 132	977 \pm 99	1209 \pm 81
5. Button-press-wall	3186 \pm 28	3143 \pm 111	3072 \pm 48	3193\pm101	553 \pm 102	667 \pm 68	615 \pm 96
6. Button-press-wall	3118 \pm 73	3344\pm108	3170 \pm 64	3275 \pm 44	495 \pm 108	711 \pm 156	633 \pm 87
7. Coffee-button	2746 \pm 53	3465\pm71	2731 \pm 336	3309 \pm 96	684 \pm 249	204 \pm 13	211 \pm 25
8. Coffee-pull	3123\pm79	1209 \pm 45	1385 \pm 128	877 \pm 118	58 \pm 4	40 \pm 2	40 \pm 4
9. Coffee-push	3197\pm39	1175 \pm 203	1463 \pm 193	729 \pm 55	228 \pm 62	41 \pm 4	65 \pm 16
10. Dial-turn	2861 \pm 70	3711\pm134	3396 \pm 108	3607 \pm 197	670 \pm 30	942 \pm 98	803 \pm 76
11. Disassemble	2990\pm85	2823 \pm 431	2811 \pm 291	2878 \pm 254	124 \pm 20	156 \pm 11	130 \pm 10
12. Door-close	2975 \pm 93	4310 \pm 44	4328 \pm 63	4481\pm19	1138 \pm 161	4359 \pm 27	2661 \pm 580
13. Door-open	3080 \pm 50	4004 \pm 116	3948 \pm 76	4010\pm109	636 \pm 61	624 \pm 76	607 \pm 62
14. Drawer-close	2936 \pm 35	4443 \pm 125	4730 \pm 28	4748\pm22	3625 \pm 462	4375 \pm 92	4482 \pm 122
15. Drawer-open	2855 \pm 112	4638 \pm 6	4065 \pm 99	4391\pm14	1746 \pm 95	1284 \pm 71	1356 \pm 127
16. Faucet-open	2945 \pm 94	4608 \pm 9	4276 \pm 158	4636\pm17	1669 \pm 47	2212 \pm 73	2584 \pm 299
17. Faucet-close	3074 \pm 23	4594\pm31	3997 \pm 173	4533 \pm 42	2349 \pm 214	2193 \pm 129	2174 \pm 185
18. Hammer	3013\pm68	516 \pm 28	1299 \pm 301	468 \pm 5	563 \pm 56	394 \pm 27	397 \pm 25
19. Handle-press-side	2949 \pm 27	4689 \pm 52	4707 \pm 29	4783\pm8	480 \pm 68	489 \pm 71	1377 \pm 270
20. Handle-press	2941 \pm 53	4648\pm66	4601 \pm 48	4579 \pm 64	2063 \pm 73	1791 \pm 100	2126 \pm 116
21. Handle-pull-side	2959 \pm 56	3647\pm143	3060 \pm 275	3340 \pm 240	165 \pm 64	19 \pm 2	24 \pm 2
22. Handle-pull	3000 \pm 83	3482 \pm 342	4019\pm59	2996 \pm 260	996 \pm 392	40 \pm 8	87 \pm 23
23. Lever-pull	2999\pm46	762 \pm 79	381 \pm 33	878 \pm 89	276 \pm 13	240 \pm 10	232 \pm 11
24. Peg-insert-side	2978\pm42	1307 \pm 81	1616 \pm 128	1238 \pm 50	192 \pm 36	11 \pm 0	10 \pm 1
25. Pick-place-wall	3102\pm60	2542 \pm 157	2728 \pm 136	1812 \pm 98	491 \pm 181	0 \pm 0	2 \pm 0
26. Pick-out-of-hole	2808\pm91	781 \pm 183	1206 \pm 142	922 \pm 122	23 \pm 5	10 \pm 1	13 \pm 1
27. Push	3157 \pm 49	2839 \pm 131	3313\pm108	2823 \pm 142	1555 \pm 125	3105 \pm 126	3193 \pm 89
28. Push-back	2897\pm111	1284 \pm 88	1799 \pm 66	881 \pm 203	16 \pm 2	7 \pm 1	5 \pm 0
29. Push	3094 \pm 115	2428 \pm 123	3421\pm56	2247 \pm 92	651 \pm 156	55 \pm 6	60 \pm 8
30. Pick-place	2921\pm54	1632 \pm 76	2165 \pm 83	1449 \pm 141	291 \pm 115	8 \pm 0	10 \pm 1
31. Plate-slide-side	2942 \pm 63	2516 \pm 138	2207 \pm 77	3221\pm98	1929 \pm 195	359 \pm 18	546 \pm 40
32. Plate-slide-side	3022 \pm 114	3250 \pm 84	2873 \pm 85	3784\pm178	818 \pm 54	191 \pm 31	662 \pm 139
33. Plate-slide back	2764 \pm 46	4235\pm28	4109 \pm 20	4165 \pm 52	1021 \pm 52	556 \pm 15	561 \pm 67
34. Plate-slide-back-side	3021 \pm 29	3776 \pm 179	4173\pm24	4124 \pm 58	631 \pm 142	183 \pm 35	728 \pm 134
35. Peg-unplug-side	2935\pm36	1390 \pm 217	1984 \pm 187	890 \pm 103	43 \pm 9	33 \pm 3	27 \pm 1
36. Soccer	2974\pm67	1056 \pm 25	1079 \pm 41	1052 \pm 65	515 \pm 152	278 \pm 30	321 \pm 16
37. Stick-push	3074\pm21	612 \pm 206	289 \pm 166	26 \pm 8	125 \pm 28	11 \pm 1	14 \pm 2
38. Stick-pull	2980\pm88	289 \pm 92	82 \pm 43	16 \pm 3	129 \pm 37	11 \pm 1	12 \pm 1
39. Push-wall	2925 \pm 74	2224 \pm 97	3586\pm50	2548 \pm 308	854 \pm 218	33 \pm 1	48 \pm 10
40. Reach-wall	3063 \pm 41	2747 \pm 263	3447\pm180	2330 \pm 273	1602 \pm 167	2906 \pm 263	3509 \pm 46
41. Shelf-place	3041\pm76	878 \pm 70	838 \pm 105	899 \pm 36	69 \pm 35	0 \pm 0	1 \pm 0
42. Sweep-into	2756 \pm 79	3962 \pm 140	4061\pm91	3095 \pm 422	879 \pm 157	238 \pm 43	207 \pm 45
43. Sweep	2902 \pm 59	2976 \pm 139	3251\pm110	1640 \pm 421	325 \pm 62	65 \pm 8	83 \pm 12
44. Window-open	2684 \pm 13	4142 \pm 51	4125 \pm 87	4225\pm31	710 \pm 62	471 \pm 30	795 \pm 116
45. Window-close	2828 \pm 112	4392 \pm 39	4397 \pm 28	4406\pm49	710 \pm 128	928 \pm 36	1152 \pm 96
Train mean	2961\pm149	2797 \pm 212	2879 \pm 201	2685 \pm 154	766 \pm 352	719 \pm 63	779 \pm 171
46. Bin-picking	2993\pm138	84 \pm 48	133 \pm 38	104 \pm 35	20 \pm 9	20 \pm 6	15 \pm 1
47. Box-close	2844\pm133	159 \pm 28	255 \pm 78	145 \pm 19	231 \pm 72	248 \pm 20	209 \pm 34
48. Hand-insert	2708\pm238	221 \pm 41	258 \pm 67	273 \pm 111	118 \pm 89	82 \pm 63	124 \pm 102
49. Door-lock	2960\pm335	2048 \pm 106	1901 \pm 747	1565 \pm 227	1659 \pm 184	1692 \pm 338	1589 \pm 374
50. Door-unlock	2962\pm51	1844 \pm 222	1952 \pm 206	1763 \pm 264	787 \pm 217	944 \pm 167	1219 \pm 339
Test mean	2912\pm105	871 \pm 117	900 \pm 226	770 \pm 140	563 \pm 60	597 \pm 154	631 \pm 203

Table 11: A list of all of the Meta-World tasks and a description of each task.

Task	Language instructions
assembly	pick up a nut and place it onto a peg
basketball	pick the basketball and place at the goal point
button-press-topdown	push the button down to the goal point
button-press-topdown-wall	bypass a wall and press a button from the top
button-press	press a button
button-press-wall	bypass a wall and press a button
coffee-button	push a button on the coffee machine
coffee-pull	place cup away
coffee-push	push cup to the goal point
dial-turn	rotate a dial 180 degrees
disassemble	pick a nut out of a peg
door-close	push the door to the goal point
door-open	pull the door to the goal point
drawer-close	push the drawer to the goal point
drawer-open	pull the drawer to the goal point
faucet-open	rotate the faucet counter-clockwise
faucet-close	rotate the faucet clockwise
hammer	push to the goal point with hammer
handle-press-side	press a handle down sideways
handle-press	press a handle down
handle-pull-side	pull a handle up sideways
handle-pull	pull a handle up
lever-pull	pull the lever to the goal point
peg-insert-side	insert the peg to the goal point
pick-place-wall	pick a puck, bypass a wall and place the puck
pick-out-of-hole	pick up a puck from a hole
reach	reach the goal point
push-back	push the puck back to the goal point
push	push the puck to the goal point
pick-place	pick the puck and place at the goal point
plate-slide	push the plate to the goal point
plate-slide-side	push the plate left to the goal point
plate-slide-back	push the plate back to the goal point
plate-slide-back-side	push the plate right to the goal point
peg-unplug-side	pull a peg sideways to the goal point
soccer	push a ball to the goal point
stick-push	grasp a stick and push a box using the stick
stick-pull	grasp a stick and pull a box with the stick
push-wall	bypass a wall and push a puck to a goal
reach-wall	bypass a wall and reach a goal
shelf-place	pick the puck and place on shelf at the goal point
sweep-into	sweep the puck into the box
sweep	sweep the puck off the table
window-open	push the window to the goal point
window-close	push the window to the goal point
bin-picking	grasp the puck from one bin and place it into another bin
box-close	grasp the cover and close the box with it
hand-insert	insert the gripper into a hole
door-lock	rotate the lock clockwise
door-unlock	rotate the lock counter-clockwise

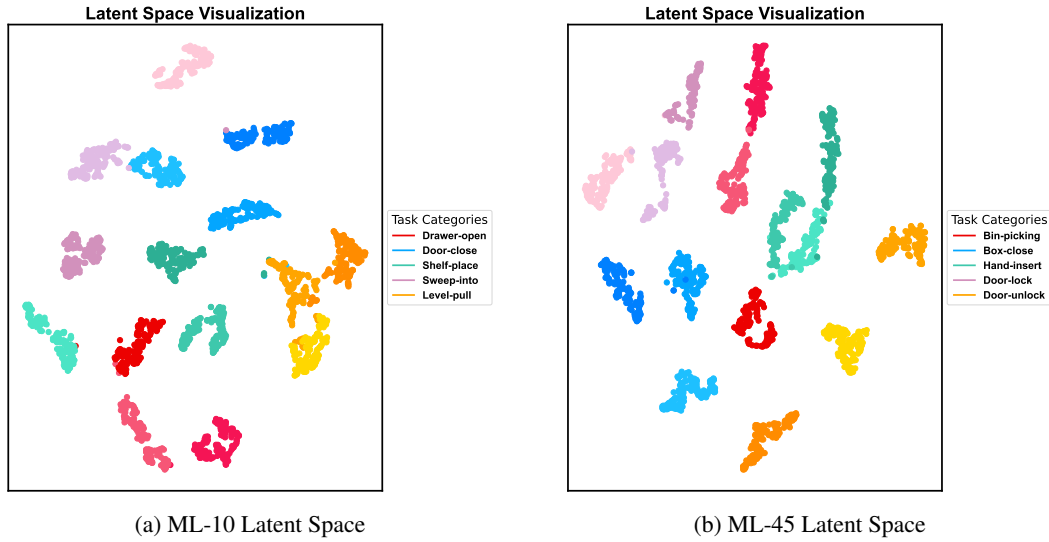


Figure 6: **Latent Space Visualization.** The t-SNE visualization of the learned task representation space for the ML-10 testing tasks is presented. We sampled three tasks from each task category of the test tasks, with each color scheme representing a different task category. Each point in the visualization corresponds to a task representation vector extracted from transitions and is color-coded according to the task properties.