A Pseudo Code

Algorithm 1 CMTA

```
Initialize: replay buffer D with ∅
Initialize: initial hidden state h_0 with zero tensor
Initialize: policy π with ϕ, Q-function Q, task encoder g, k experts f^1, ···, f^k, lstm, fully connected layer W
Input: state s_t for each environment, one-hot task id z_τ
for episode m = 1, 2, ··· do
    for time-step t = 1, 2, ··· do
        for each task τ_i do
            z_enc^j = f^j(s_t), ∀j ∈ 1, ···, k
            z_task = g(z_τ)
            h_t = lstm(s_t; h_{t−1})
            α_1, ···, α_k = softmax(W(h_t; z_task))
            z_enc = Σ_{j=1}^{k} α_j · z_enc^j
            z = z_task || z_enc
            sample action a_t ∼ π(·|z_task; z_enc)
            Perform action a_t, get reward r_t and next state s_{t+1}.
            D = D ⋃ {s_t, a_t, r_t, s_{t+1}, h_t, z_τ}
        end for
        randomly sample batch from D
        compute L_contrastive by Eq 3
        compute L_actor and L_critic by Eq9 and Eq10
        update k experts with L_contrastive
        update π_ϕ with L_actor
        update all components except π_ϕ with L_critic
    end for
end for
```

B Libraries

We use the following open-source libraries: MetaWorld[2], MTEnv[3],MTRL[4].

[2]https://github.com/rlworkgroup/metaworld, commit-id:af8417bfc82a3e249b4b02156518d775f29eb289

[3]https://github.com/facebookresearch/ mtenv

[4]https://github.com/facebookresearch/mtrl

C Additional Experiment Results

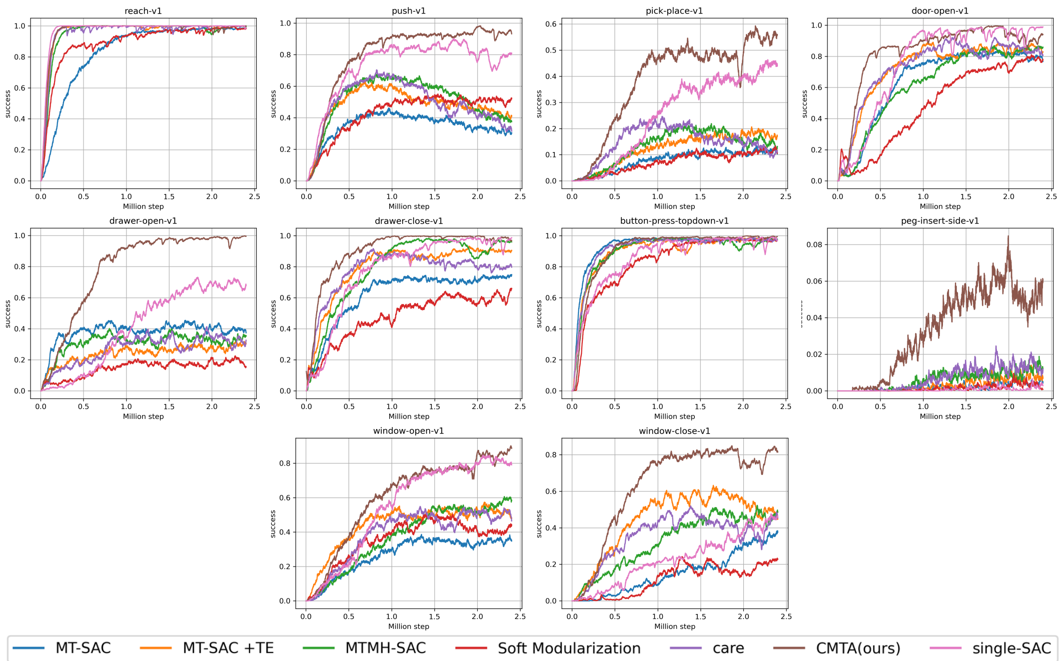

Figure 6: Training curves of different methods on each task of MT10-Mixed, each curve is averaged over 8 seeds. Our approach consistently outperforms baselines in all tasks, whether on asymptotic performance or sample efficiency.

D Hyperparameter Details.

Table 3: Hyperparameter values that are common across all the methods

Hyperparameter	Hyperparameter values
batch size	128 × number of tasks
network architecture	feedforward network
actor/critic size	three fully connected layers with 512 units
non-linearity	ReLU
policy initialization	standard Gaussian
temperature	learned and distangled with tasks
exploration parameters	run a uniform exploration policy 1500 steps
num of samples / num of train steps per iteration	1 env step / 1 training step
evaluation frequency	3000 steps
replay buffer size	5000000
policy learning rate	3e-4
Q function learning rate	3e-4
optimizer	Adam
policy learning rate	3e-4
beta for Adam optimizer for policy	(0.9, 0.999)
Q function learning rate	3e-4
beta for Adam optimizer for Q function	(0.9, 0.999)
discount	0.99
Episode length (horizon)	150
reward scale	1

Table 4: Hyperparameter values of task encoder

Hyperparameter	Hyperparameter values
task encoder train from scratch	embedding layer with dim 64 + FC 128 + FC 64 + FC 64
pretrained	pre-trained embedding layer with dim 512 + FC 128 + FC 64 + FC 64

Table 5: Hyperparameter values of Soft Modularization

Hyperparameter	Hyperparameter values
task encoder type	train from scratch
routing network size	4 layers and 4 modules per layer with dim 64

Table 6: Hyperparameter values of CARE

Hyperparameter	Hyperparameter values
task encoder type	pre-trained embedding layer
encoder size	FC 64 + FC 64
number of encoders	6

Table 7: Hyperparameter values of CMTA

Hyperparameter	Hyperparameter values
task encoder type	train from scratch
encoder size	FC 64 + FC 64
number of encoders	6
β	2500