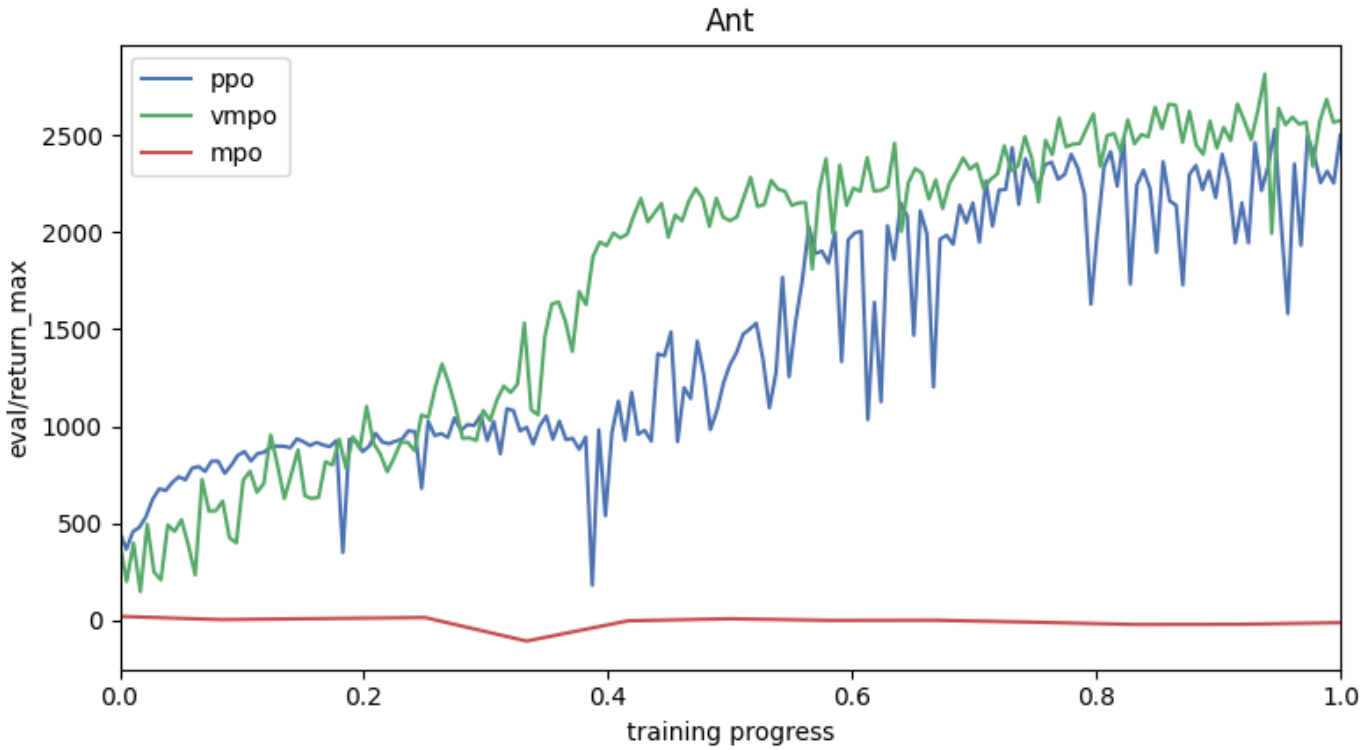
Contents

Report

Environment	ppo	vmpo	mpo	vmpo_sgd	ppo_lm	nanochat_rl
Ant	2499	2574	-13	-	-	-
HalfCheetah	1793	3398	3656	246	-	-
Humanoid	401	723	5039	298	-	-
Walker2d	251	727	2848	-	-	-
dm_control/cartpole/swingup	296	335	-	-	-	-
dm_control/cheetah/run	449	393	323	174	-	-
dm_control/humanoid/walk	14	4	8	-	-	-
dm_control/walker/run	412	132	597	-	-	-
dm_control/walker/walk	666	281	969	73	-	-

Ant



ppo config:

```
{
  "clip_ratio": 0.15,
```

```
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256
  ],
  "ent_coef": 0.0003,
  "env": "Ant-v5",
  "eval_interval": 10000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 512,
  "normalize_obs": true,
  "num_envs": 16,
  "out_dir": "checkpoints/ppo/Ant-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "rollout_steps": 8192,
  "save_interval": 1000000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 30000000,
  "update_epochs": 12,
  "value_lr": 2e-05,
  "vf_coef": 0.5,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo",
  "env": "Ant-v5",
  "epsilon_eta": 0.7,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.001,
  "eval_interval": 20000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/Ant-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 200000,
  "seed": 42,
  "temperature_init": 2,
```

```
  "temperature_lr": 0.001,
  "topk_fraction": 0.2,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0003,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```

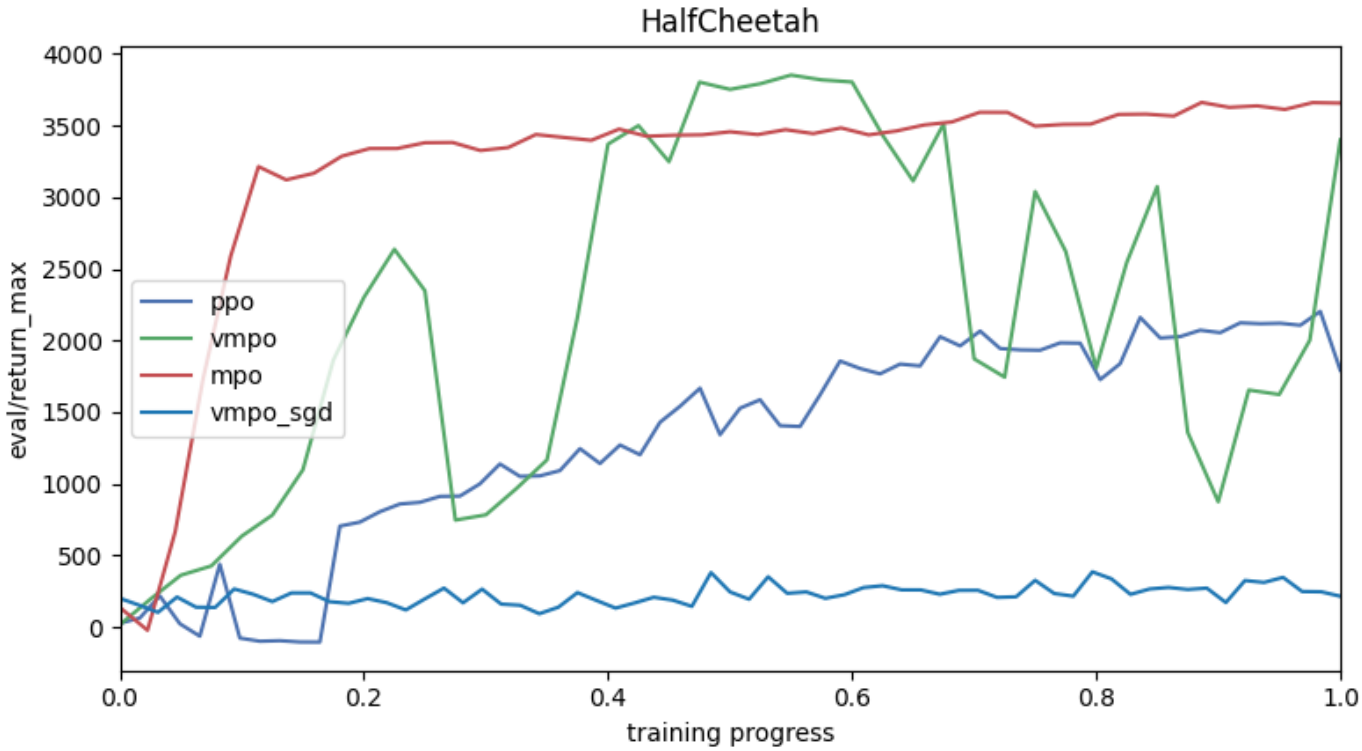

mpo config:


```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 512,
  "command": "mpo",
  "critic_layer_sizes": [
    512,
    512,
    256
  ],
  "env": "Ant-v5",
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
  "epsilon_stddev": null,
  "eval_interval": 7000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/Ant-v5",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
  "total_steps": 50000000,
  "update_after": 10000,
  "updates_per_step": 1,
  "use_retrace": true,
  "wandb_entity": null,
```

```
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-Ant-v5	mpo	98718	-13
mpo-Ant-v5	mpo	86448	-22
ppo-Ant-v5	ppo	30000000	2499
vmpo-Ant-v5	vmpo	3583657	2574
vmpo-Ant-v5	vmpo	432719	659
vmpo-Ant-v5	vmpo	233862	386

HalfCheetah



ppo config:


```
{
  "clip_ratio": 0.15,
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256
  ],
  "ent_coef": 0.0003,
  "env": "HalfCheetah-v5",
  "eval_interval": 10000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 512,
  "normalize_obs": true,
  "num_envs": 16,
  "out_dir": "checkpoints/ppo/HalfCheetah-v5",
```

```
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "rollout_steps": 8192,
  "save_interval": 1000000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 10000000,
  "update_epochs": 12,
  "value_lr": 2e-05,
  "vf_coef": 0.5,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo",
  "env": "HalfCheetah-v5",
  "epsilon_eta": 0.7,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.001,
  "eval_interval": 20000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/HalfCheetah-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 200000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.001,
  "topk_fraction": 0.2,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0003,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


mpo config:

```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 512,
  "command": "mpo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "env": "HalfCheetah-v5",
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
  "epsilon_stddev": null,
  "eval_interval": 7000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/HalfCheetah-v5",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
  "total_steps": 50000000,
  "update_after": 10000,
  "updates_per_step": 1,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo_sgd config:

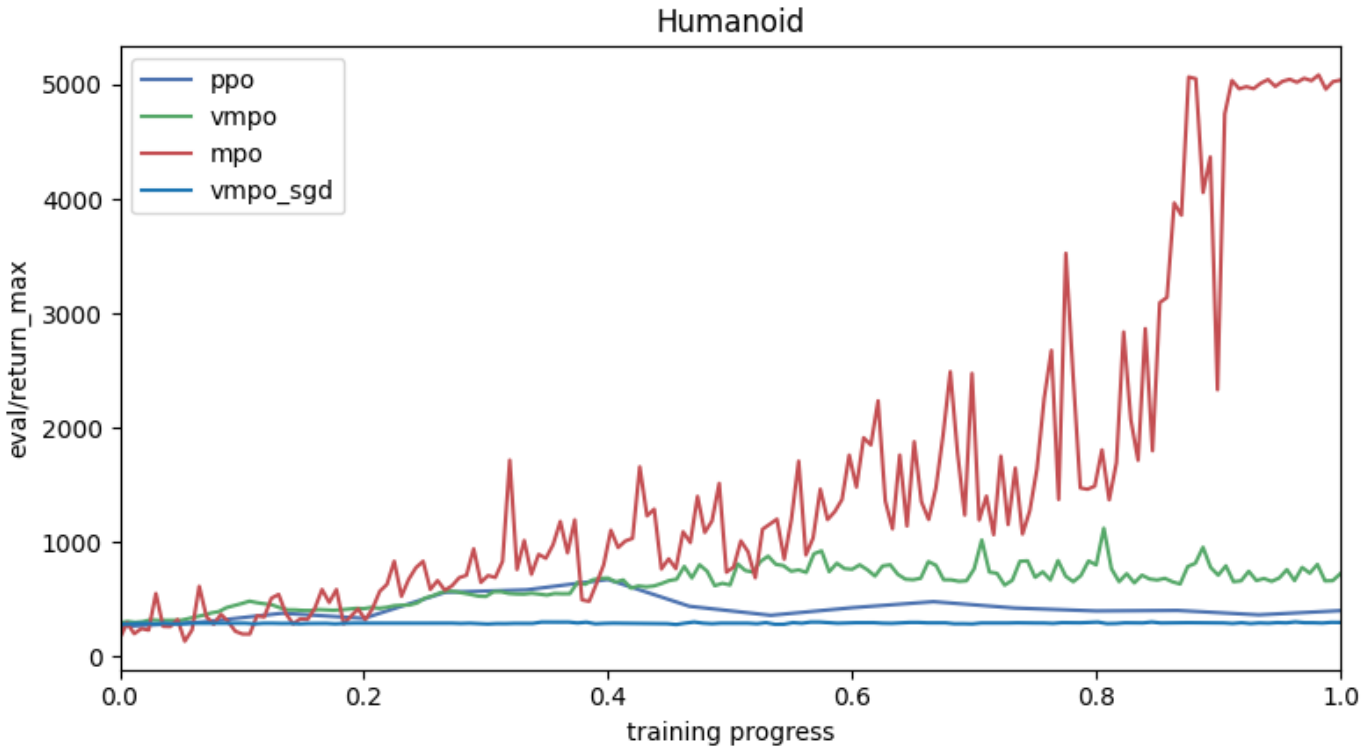

```
{
  "alpha_lr": 0.0001,
  "command": "vmpo_sgd",
  "device": null,
  "env": "HalfCheetah-v5",
  "epsilon_eta": 0.7,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.001,
  "eval_interval": 20000,
  "gamma": 0.99,
```

```
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "optimizer_type": "sgd",
  "out_dir": "checkpoints/vmpo_sgd/HalfCheetah-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 200000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.001,
  "topk_fraction": 0.2,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0003,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-HalfCheetah-v5	mpo	327895	3656
mpo-HalfCheetah-v5	mpo	596969	3284
ppo-HalfCheetah-v5	ppo	10000000	1793
ppo-HalfCheetah-v5	ppo	3679985	1158
ppo-HalfCheetah-v5	ppo	3087985	-10
ppo-HalfCheetah-v5	ppo	393216	-193
ppo-HalfCheetah-v5	ppo	262144	-221
vmpo-HalfCheetah-v5	vmpo	830829	3398
vmpo-HalfCheetah-v5	vmpo	4840000	435
vmpo-HalfCheetah-v5	vmpo	1126124	299
vmpo-HalfCheetah-v5	vmpo	240000	213
vmpo-HalfCheetah-v5	vmpo	57344	210
vmpo-HalfCheetah-v5	vmpo	180179	196
vmpo-HalfCheetah-v5	vmpo	284672	164
vmpo-HalfCheetah-v5	vmpo	581632	76
vmpo-HalfCheetah-v5	vmpo	81000	33
vmpo-HalfCheetah-v5	vmpo	233472	20
vmpo-HalfCheetah-v5	vmpo	946945	-166
vmpo-HalfCheetah-v5	vmpo	593592	-279
vmpo-HalfCheetah-v5	vmpo	1000999	-299
vmpo-HalfCheetah-v5	vmpo	526525	-299
vmpo-HalfCheetah-v5	vmpo	230229	-345
vmpo-HalfCheetah-v5	vmpo	803802	-349
vmpo-HalfCheetah-v5	vmpo	485484	-451
vmpo-HalfCheetah-v5	vmpo	51200	-1247
vmpo-HalfCheetah-v5	vmpo	299298	-

Run	Algorithm	_step	eval/return_max
vmpo-HalfCheetah-v5	vmpo	197196	-
vmpo-HalfCheetah-v5	vmpo	258048	-
vmpo_sgd-HalfCheetah-v5	vmpo_sgd	1296000	246
vmpo_sgd-HalfCheetah-v5	vmpo_sgd	1162000	110

Humanoid



ppo config:


```
{
  "clip_ratio": 0.15,
  "command": "ppo",
  "critic_layer_sizes": [
    512,
    256,
    256
  ],
  "ent_coef": 0.0003,
  "env": "Humanoid-v5",
  "eval_interval": 25000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 512,
  "normalize_obs": true,
  "num_envs": 8,
  "out_dir": "checkpoints/ppo/Humanoid-v5",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
```

```
  "policy_lr": 0.0001,
  "rollout_steps": 8192,
  "save_interval": 1000000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 10000000,
  "update_epochs": 12,
  "value_lr": 2e-05,
  "vf_coef": 0.5,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo",
  "env": "Humanoid-v5",
  "epsilon_eta": 0.25,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.0003,
  "eval_interval": 10000,
  "gamma": 0.995,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/Humanoid-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 5e-05,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 100000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.0005,
  "topk_fraction": 0.1,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0001,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


mpo config:


```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 512,
```

```
  "command": "mpo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "env": "Humanoid-v5",
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
  "epsilon_stddev": null,
  "eval_interval": 2000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/Humanoid-v5",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
  "total_steps": 50000000,
  "update_after": 500000,
  "updates_per_step": 1,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```

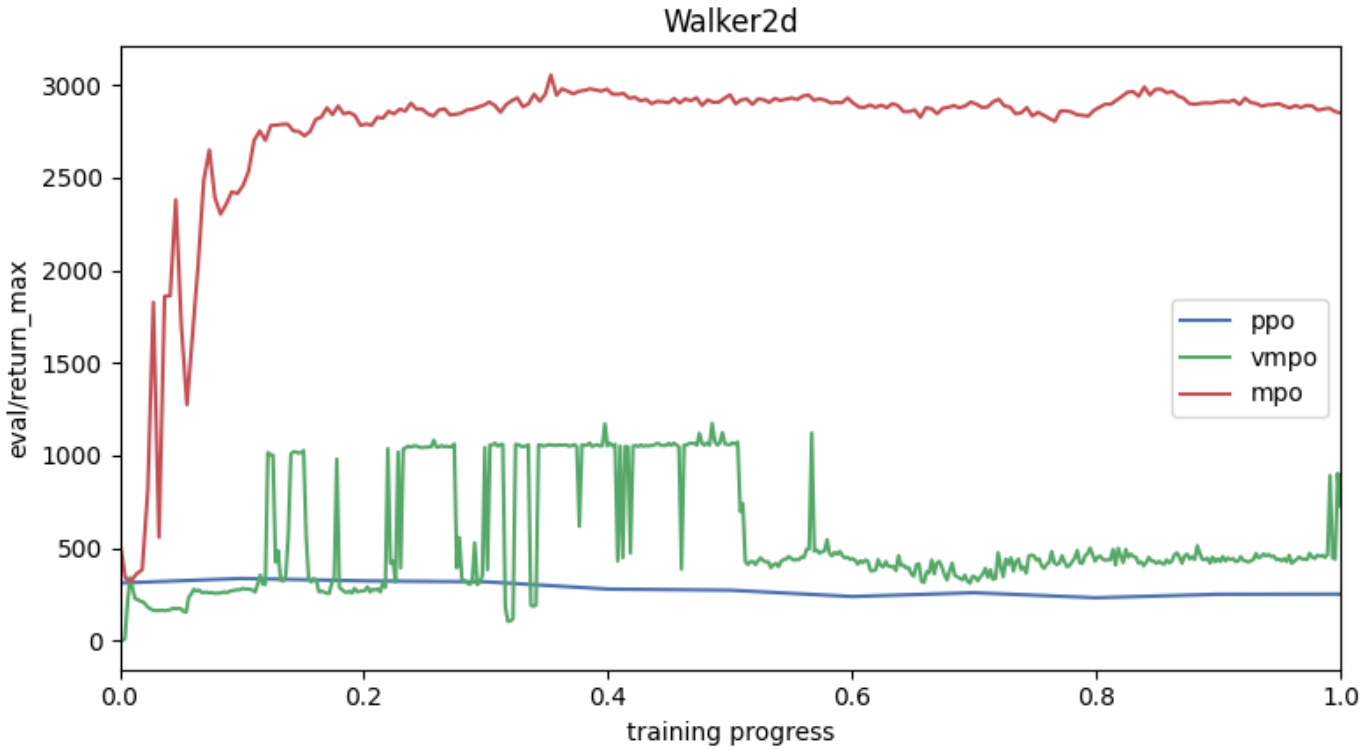

vmpo_sgd config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo_sgd",
  "device": null,
  "env": "Humanoid-v5",
  "epsilon_eta": 0.25,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.0003,
  "eval_interval": 10000,
  "gamma": 0.995,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "optimizer_type": "sgd",
```

```
  "out_dir": "checkpoints/vmpo_sgd/Humanoid-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 5e-05,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 100000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.0005,
  "topk_fraction": 0.1,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0001,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-Humanoid-v5	mpo	839364	5039
ppo-Humanoid-v5	ppo	3342336	401
ppo-Humanoid-v5	ppo	309657	260
vmpo-Humanoid-v5	vmpo	1614700	723
vmpo-Humanoid-v5	vmpo	349790	599
vmpo-Humanoid-v5	vmpo	643914	556
vmpo-Humanoid-v5	vmpo	4489204	518
vmpo-Humanoid-v5	vmpo	312590	415
vmpo-Humanoid-v5	vmpo	1278677	292
vmpo-Humanoid-v5	vmpo	153081	251
vmpo-Humanoid-v5	vmpo	258347	170
vmpo-Humanoid-v5	vmpo	107190	130
vmpo-Humanoid-v5	vmpo	294912	123
vmpo-Humanoid-v5	vmpo	257922	115
vmpo-Humanoid-v5	vmpo	166628	105
vmpo-Humanoid-v5	vmpo	180221	105
vmpo-Humanoid-v5	vmpo	319248	104
vmpo-Humanoid-v5	vmpo	128475	101
vmpo-Humanoid-v5	vmpo	399869	74
vmpo-Humanoid-v5	vmpo	1047804	66
vmpo-Humanoid-v5	vmpo	118456	-
vmpo-Humanoid-v5	vmpo	85334	-
vmpo-Humanoid-v5	vmpo	129131	-
vmpo-Humanoid-v5	vmpo	159994	-
vmpo_sgd-Humanoid-v5	vmpo_sgd	1375920	298
vmpo_sgd-Humanoid-v5	vmpo_sgd	2032252	285

Walker2d



ppo config:


```
{
  "clip_ratio": 0.15,
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256
  ],
  "ent_coef": 0.0003,
  "env": "Walker2d-v5",
  "eval_interval": 10000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 512,
  "normalize_obs": true,
  "num_envs": 16,
  "out_dir": "checkpoints/ppo/Walker2d-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "rollout_steps": 8192,
  "save_interval": 1000000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 30000000,
  "update_epochs": 12,
  "value_lr": 2e-05,
  "vf_coef": 0.5,
```

```
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo",
  "env": "Walker2d-v5",
  "epsilon_eta": 0.7,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.001,
  "eval_interval": 20000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/Walker2d-v5",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 200000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.001,
  "topk_fraction": 0.2,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0003,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


mpo config:


```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 512,
  "command": "mpo",
  "critic_layer_sizes": [
    512,
    256,
    256
  ],
  "env": "Walker2d-v5",
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
```

```
  "epsilon_stddev": null,
  "eval_interval": 7000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/Walker2d-v5",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
  "total_steps": 50000000,
  "update_after": 10000,
  "updates_per_step": 1,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-Walker2d-v5	mpo	1541254	2848
ppo-Walker2d-v5	ppo	1915489	251
vmpo-Walker2d-v5	vmpo	9589342	727

dm_control/cartpole/swingup

ppo config:


```
{
  "clip_ratio": 0.2,
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "ent_coef": 0.001,
  "env": "dm_control/cartpole/swingup",
  "eval_interval": 50000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 128,
  "normalize_obs": false,
  "num_envs": 1,
  "out_dir": "checkpoints/ppo/dm_control-cartpole-swingup",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "rollout_steps": 512,
  "save_interval": 1000000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 1000000,
  "update_epochs": 4,
```

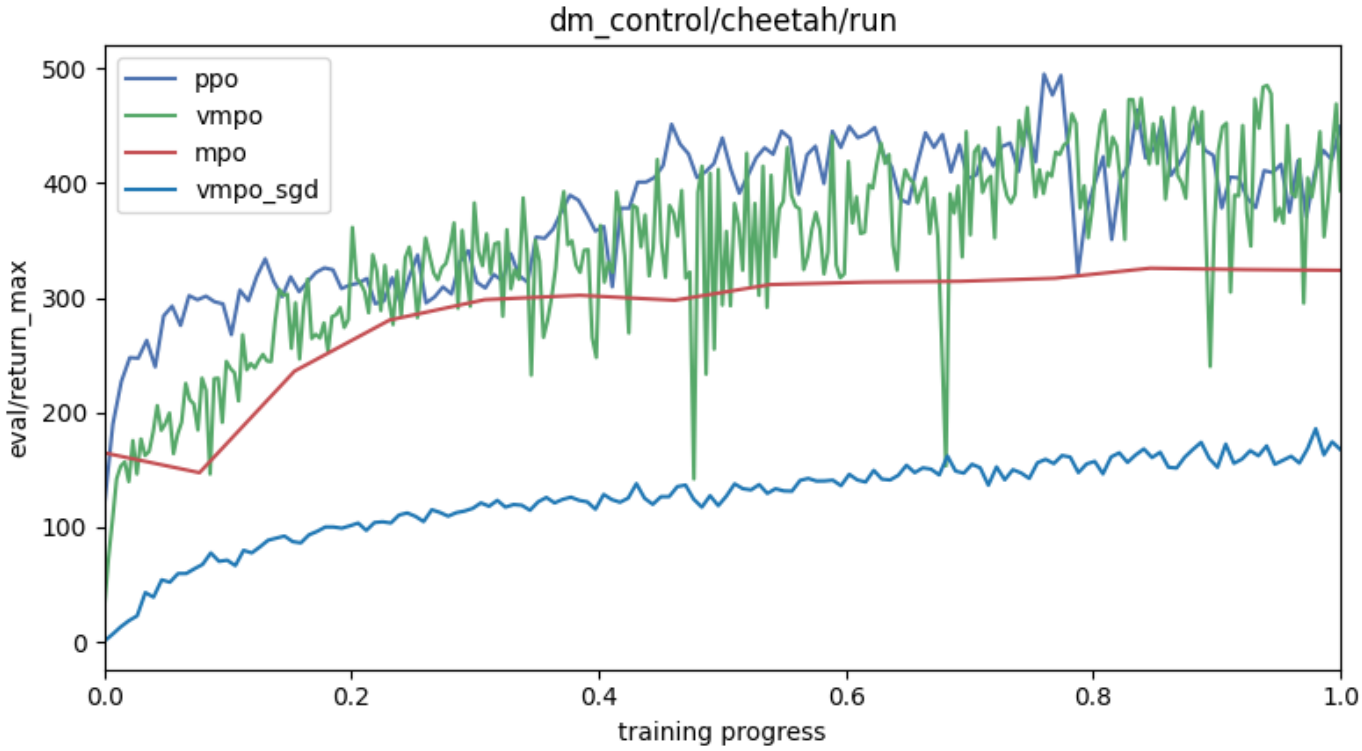
```
  "value_lr": 0.0001,
  "vf_coef": 0.5,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.001,
  "command": "vmpo",
  "env": "dm_control/cartpole/swingup",
  "epsilon_eta": 0.1,
  "epsilon_mu": 0.01,
  "epsilon_sigma": 0.0001,
  "eval_interval": 10000,
  "gamma": 0.99,
  "kl_mean_coef": 0.001,
  "kl_std_coef": 0.001,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 1,
  "out_dir": "checkpoints/vmpo/dm_control-cartpole-swingup",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0001,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 2048,
  "save_interval": 25000,
  "seed": 42,
  "temperature_init": 5,
  "temperature_lr": 0.001,
  "topk_fraction": 1,
  "total_steps": 2000000,
  "updates_per_step": 10,
  "value_layer_sizes": [
    256,
    256
  ],
  "value_lr": 0.0001,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
ppo-dm_control/cartpole/swingup	ppo	1000000	296
ppo-dm_control/cartpole/swingup	ppo	195194	159
ppo-dm_control/cartpole/swingup	ppo	915456	130
vmpo-dm_control/cartpole/swingup	vmpo	2000000	335

dm_control/cheetah/run



ppo config:


```
{
  "clip_ratio": 0.25,
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "ent_coef": 0.0001,
  "env": "dm_control/cheetah/run",
  "eval_interval": 10000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 256,
  "normalize_obs": true,
  "num_envs": 1,
  "out_dir": "checkpoints/ppo/dm_control-cheetah-run",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0002,
  "rollout_steps": 2048,
  "save_interval": 1000000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 3000000,
  "update_epochs": 4,
```

```
  "value_lr": 0.0003,
  "vf_coef": 1,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo",
  "env": "dm_control/cheetah/run",
  "epsilon_eta": 0.1,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.0006,
  "eval_interval": 25000,
  "gamma": 0.99,
  "max_grad_norm": 2,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/dm_control-cheetah-run",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0002,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 500000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.0003,
  "topk_fraction": 0.25,
  "total_steps": 10000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 5e-05,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


mpo config:


```
{
  "action_penalization": false,
  "action_samples": 128,
  "batch_size": 256,
  "command": "mpo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "env": "dm_control/cheetah/run",
```

```
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
  "epsilon_stddev": null,
  "eval_interval": 10000,
  "gamma": 0.995,
  "kl_epsilon": 0.2,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/dm_control-cheetah-run",
  "per_dim_constraining": true,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 1,
  "temperature_lr": 0.0003,
  "total_steps": 20000000,
  "update_after": 1000,
  "updates_per_step": 1,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```

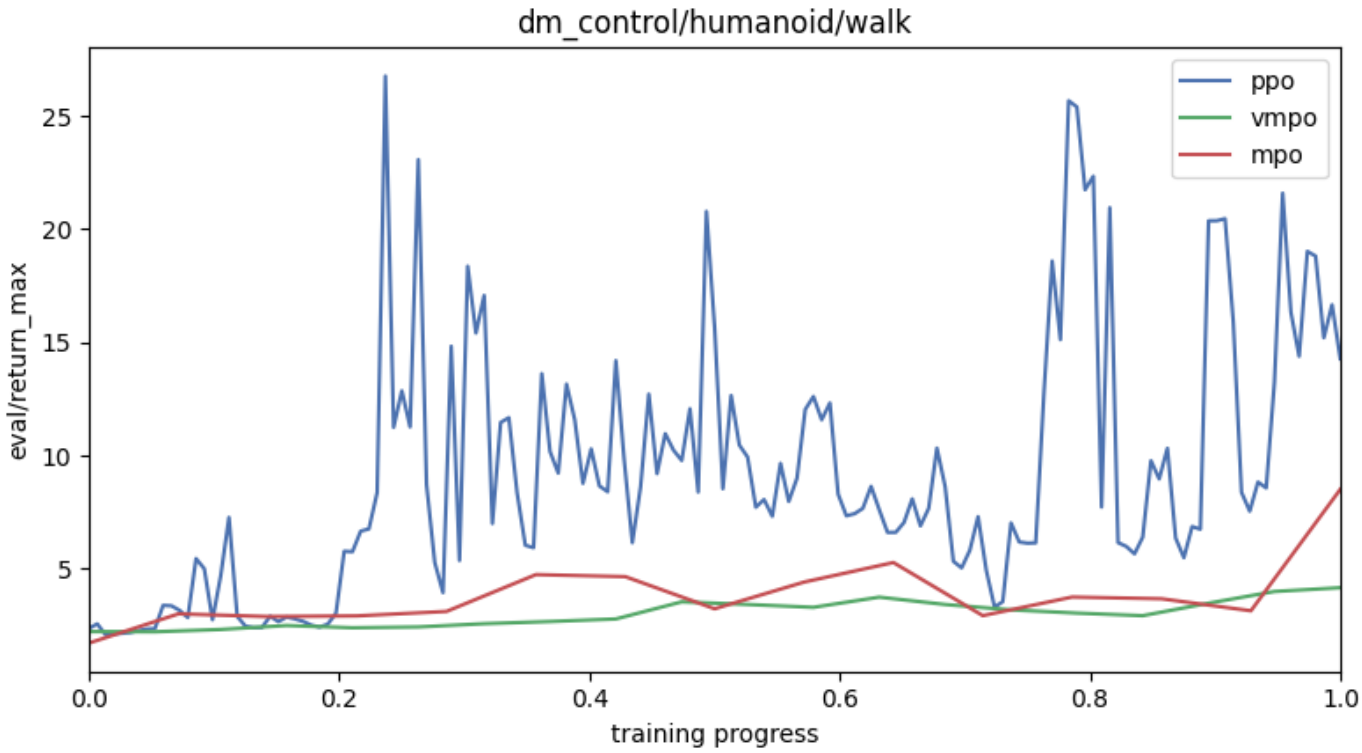

vmpo_sgd config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo_sgd",
  "device": null,
  "env": "dm_control/cheetah/run",
  "epsilon_eta": 0.1,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.0006,
  "eval_interval": 25000,
  "gamma": 0.99,
  "max_grad_norm": 2,
  "normalize_advantages": true,
  "num_envs": 16,
  "optimizer_type": "sgd",
  "out_dir": "checkpoints/vmpo_sgd/dm_control-cheetah-run",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0002,
  "popart_beta": 0.0001,
```

```
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 500000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.0003,
  "topk_fraction": 0.25,
  "total_steps": 10000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 5e-05,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-dm_control/cheetah/run	mpo	145824	323
ppo-dm_control/cheetah/run	ppo	1474560	449
ppo-dm_control/cheetah/run	ppo	1605632	313
ppo-dm_control/cheetah/run	ppo	120113	260
vmpo-dm_control/cheetah/run	vmpo	7642634	393
vmpo-dm_control/cheetah/run	vmpo	790000	219
vmpo-dm_control/cheetah/run	vmpo	1572000	191
vmpo-dm_control/cheetah/run	vmpo	452451	191
vmpo-dm_control/cheetah/run	vmpo	250000	131
vmpo-dm_control/cheetah/run	vmpo	454453	87
vmpo_sgd-dm_control/cheetah/run	vmpo_sgd	3799000	174
vmpo_sgd-dm_control/cheetah/run	vmpo_sgd	2648000	66

dm_control/humanoid/walk



ppo config:


```
{
  "clip_ratio": 0.2,
  "command": "ppo",
  "critic_layer_sizes": [
    512,
    256
  ],
  "ent_coef": 0.0005,
  "env": "dm_control/humanoid/walk",
  "eval_interval": 2000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 256,
  "normalize_obs": true,
  "num_envs": 16,
  "out_dir": "checkpoints/ppo/dm_control-humanoid-walk",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0002,
  "rollout_steps": 2048,
  "save_interval": 250000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 5000000,
  "update_epochs": 2,
  "value_lr": 5e-05,
  "vf_coef": 0.5,
```

```
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0001,
  "command": "vmpo",
  "env": "dm_control/humanoid/walk",
  "epsilon_eta": 0.7,
  "epsilon_mu": 0.05,
  "epsilon_sigma": 0.001,
  "eval_interval": 20000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/dm_control-humanoid-walk",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0001,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 200000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.001,
  "topk_fraction": 0.2,
  "total_steps": 30000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0003,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```

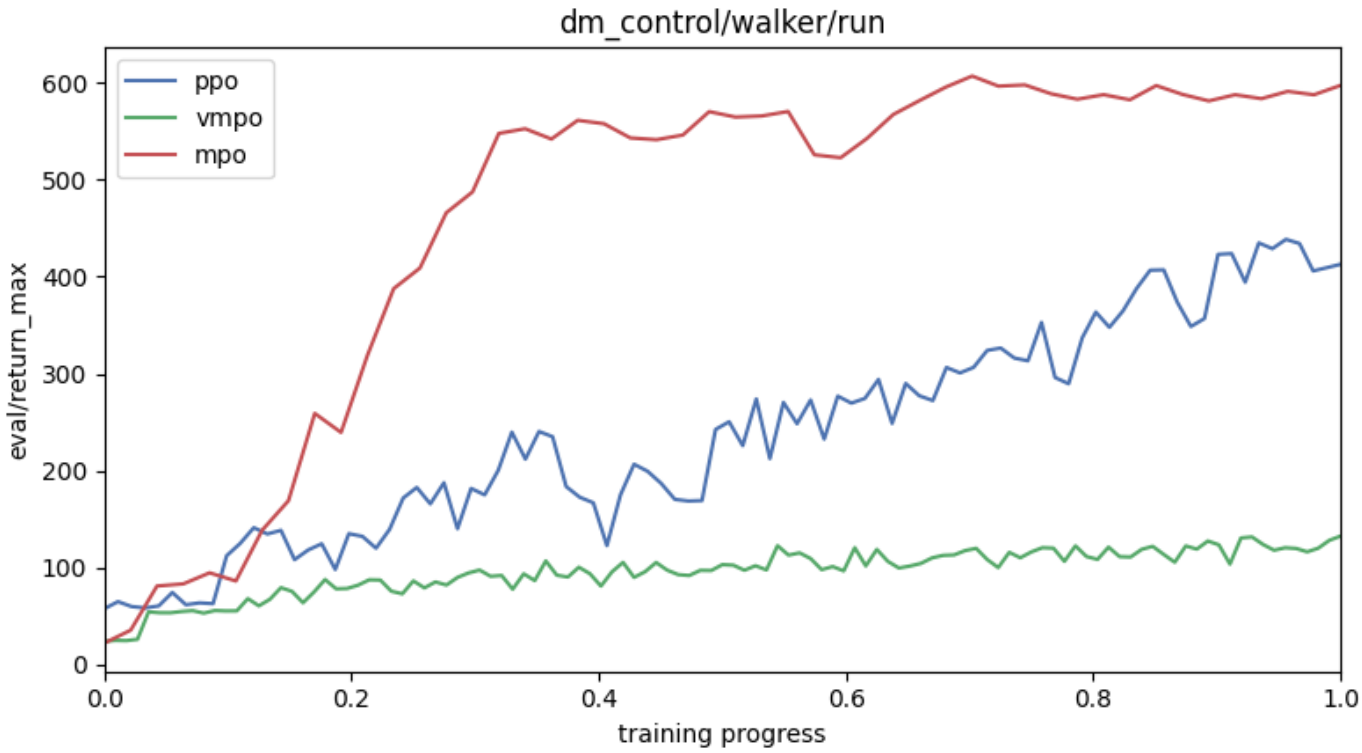

mpo config:


```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 512,
  "command": "mpo",
  "critic_layer_sizes": [
    512,
    256,
    256
  ],
  "env": "dm_control/humanoid/walk",
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
```

```
  "epsilon_stddev": null,
  "eval_interval": 20000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/dm_control-humanoid-walk",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 1,
  "temperature_lr": 0.0003,
  "total_steps": 50000000,
  "update_after": 500000,
  "updates_per_step": 2,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-dm_control/humanoid/walk	mpo	794816	8
mpo-dm_control/humanoid/walk	mpo	683232	4
ppo-dm_control/humanoid/walk	ppo	5000000	14
vmpo-dm_control/humanoid/walk	vmpo	412000	4
vmpo-dm_control/humanoid/walk	vmpo	5359000	2
vmpo-dm_control/humanoid/walk	vmpo	891890	2

dm_control/walker/run



ppo config:


```
{
  "clip_ratio": 0.2,
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "ent_coef": 0.0001,
  "env": "dm_control/walker/run",
  "eval_interval": 15000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 256,
  "normalize_obs": true,
  "num_envs": 4,
  "out_dir": "checkpoints/ppo/dm_control-walker-run",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0002,
  "rollout_steps": 2048,
  "save_interval": 50000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 20000000,
  "update_epochs": 4,
```

```
  "value_lr": 0.0001,
  "vf_coef": 0.5,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0005,
  "command": "vmpo",
  "env": "dm_control/walker/run",
  "epsilon_eta": 0.05,
  "epsilon_mu": 0.01,
  "epsilon_sigma": 0.0001,
  "eval_interval": 25000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "out_dir": "checkpoints/vmpo/dm_control-walker-run",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0003,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 8192,
  "save_interval": 50000,
  "seed": 42,
  "temperature_init": 2,
  "temperature_lr": 0.0003,
  "topk_fraction": 0.2,
  "total_steps": 40000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0001,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```

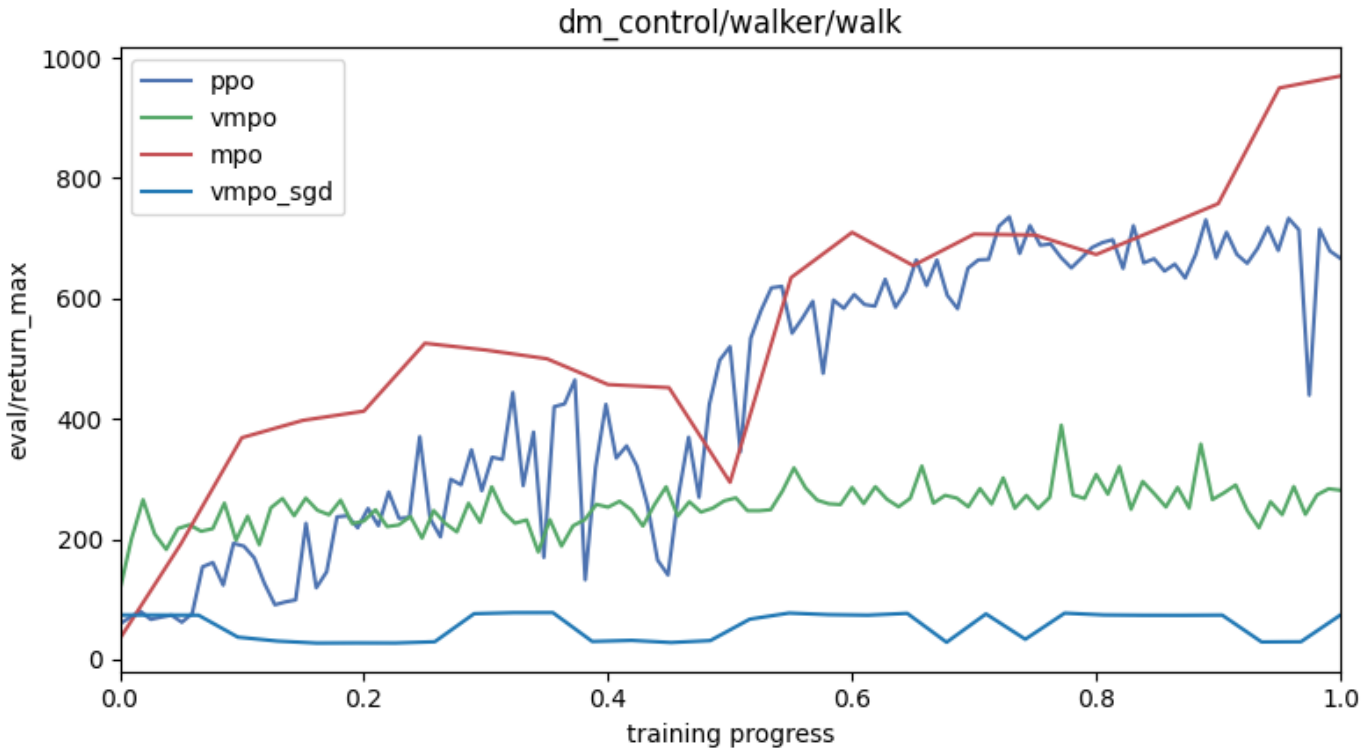

mpo config:


```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 256,
  "command": "mpo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "env": "dm_control/walker/run",
```

```
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
  "epsilon_stddev": null,
  "eval_interval": 3000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/dm_control-walker-run",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
  "total_steps": 40000000,
  "update_after": 500000,
  "updates_per_step": 2,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-dm_control/walker/run	mpo	643047	597
ppo-dm_control/walker/run	ppo	5581569	412
vmpo-dm_control/walker/run	vmpo	2834432	132
vmpo-dm_control/walker/run	vmpo	5115109	128
vmpo-dm_control/walker/run	vmpo	586000	61
vmpo-dm_control/walker/run	vmpo	5535000	49

dm_control/walker/walk



ppo config:


```
{
  "clip_ratio": 0.2,
  "command": "ppo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "ent_coef": 0.0001,
  "env": "dm_control/walker/walk",
  "eval_interval": 15000,
  "gae_lambda": 0.95,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "minibatch_size": 128,
  "normalize_obs": true,
  "num_envs": 4,
  "out_dir": "checkpoints/ppo/dm_control-walker-walk",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0002,
  "rollout_steps": 2048,
  "save_interval": 50000,
  "seed": 42,
  "target_kl": 0.02,
  "total_steps": 10000000,
  "update_epochs": 4,
```

```
  "value_lr": 0.0001,
  "vf_coef": 0.5,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo config:


```
{
  "alpha_lr": 0.0005,
  "command": "vmpo",
  "env": "dm_control/walker/walk",
  "epsilon_eta": 0.05,
  "epsilon_mu": 0.01,
  "epsilon_sigma": 0.0001,
  "eval_interval": 25000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "num_envs": 2,
  "out_dir": "checkpoints/vmpo/dm_control-walker-walk",
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "rollout_steps": 4096,
  "save_interval": 50000,
  "seed": 42,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
  "topk_fraction": 0.25,
  "total_steps": 40000000,
  "updates_per_step": 2,
  "value_lr": 0.0001,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


mpo config:


```
{
  "action_penalization": false,
  "action_samples": 256,
  "batch_size": 256,
  "command": "mpo",
  "critic_layer_sizes": [
    256,
    256,
    256
  ],
  "env": "dm_control/walker/walk",
  "epsilon_mean": null,
  "epsilon_penalty": 0.001,
  "epsilon_stddev": null,
  "eval_interval": 3000,
  "gamma": 0.995,
  "kl_epsilon": 0.1,
  "lambda_init": 1,
```

```
  "lambda_lr": 0.0003,
  "max_grad_norm": 1,
  "mstep_kl_epsilon": 0.1,
  "out_dir": "checkpoints/mpo/dm_control-walker-walk",
  "per_dim_constraining": false,
  "policy_layer_sizes": [
    256,
    256,
    256
  ],
  "policy_lr": 0.0003,
  "q_lr": 0.0003,
  "replay_size": 1000000,
  "retrace_lambda": 0.95,
  "retrace_mc_actions": 8,
  "retrace_steps": 2,
  "save_interval": 50000,
  "seed": 42,
  "tau": 0.005,
  "temperature_init": 1,
  "temperature_lr": 0.0003,
  "total_steps": 40000000,
  "update_after": 500000,
  "updates_per_step": 2,
  "use_retrace": true,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


vmpo_sgd config:


```
{
  "alpha_lr": 0.0005,
  "command": "vmpo_sgd",
  "device": null,
  "env": "dm_control/walker/walk",
  "epsilon_eta": 0.05,
  "epsilon_mu": 0.01,
  "epsilon_sigma": 0.0001,
  "eval_interval": 25000,
  "gamma": 0.99,
  "max_grad_norm": 0.5,
  "normalize_advantages": true,
  "num_envs": 16,
  "optimizer_type": "sgd",
  "out_dir": "checkpoints/vmpo_sgd/dm_control-walker-walk",
  "policy_layer_sizes": [
    256,
    256
  ],
  "policy_lr": 0.0003,
  "popart_beta": 0.0001,
  "popart_eps": 1e-08,
  "popart_min_sigma": 0.001,
  "rollout_steps": 4096,
  "save_interval": 50000,
  "seed": 42,
  "temperature_init": 3,
  "temperature_lr": 0.0003,
```

```
  "topk_fraction": 0.25,
  "total_steps": 40000000,
  "updates_per_step": 2,
  "value_layer_sizes": [
    512,
    256
  ],
  "value_lr": 0.0001,
  "wandb_entity": null,
  "wandb_group": null,
  "wandb_project": null
}
```


Run	Algorithm	_step	eval/return_max
mpo-dm_control/walker/walk	mpo	563447	969
ppo-dm_control/walker/walk	ppo	7143424	666
vmpo-dm_control/walker/walk	vmpo	5300294	281
vmpo-dm_control/walker/walk	vmpo	1638400	230
vmpo_sgd-dm_control/walker/walk	vmpo_sgd	802000	73
vmpo_sgd-dm_control/walker/walk	vmpo_sgd	944000	32