Tests_DQN

November 14, 2019

[1]:
```
import math
import random
import numpy as np
from collections import namedtuple
from itertools import count
import matplotlib.pyplot as plt

import pdb

import torch
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F
import torchvision.transforms as T
```

1 Custom Anti Collision Tests openai gym environement

[2]:
```
import gym
import gym_act

# Our Anti Collision Tests environement
env = gym.make("Act-v2")
```

```
ACT (Anti Collision Tests) with 10 cars using cv driver model
SEED 15438151168752304802
```

[3]:
```
# if gpu is to be used
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

def show_img(img):
    plt.imshow(img)
    plt.show()
```

2 Replay Buffer

[4]:
```
class ReplayMemory(object):

    def __init__(self, capacity):
        self.capacity = capacity
        self.memory = []
        self.position = 0

    def push(self, *args):
        """Saves a transition."""
        if len(self.memory) < self.capacity:
            self.memory.append(None)
        self.memory[self.position] = Transition(*args)
        self.position = (self.position + 1) % self.capacity

    def sample(self, batch_size):
        return random.sample(self.memory, batch_size)

    def __len__(self):
        return len(self.memory)
```

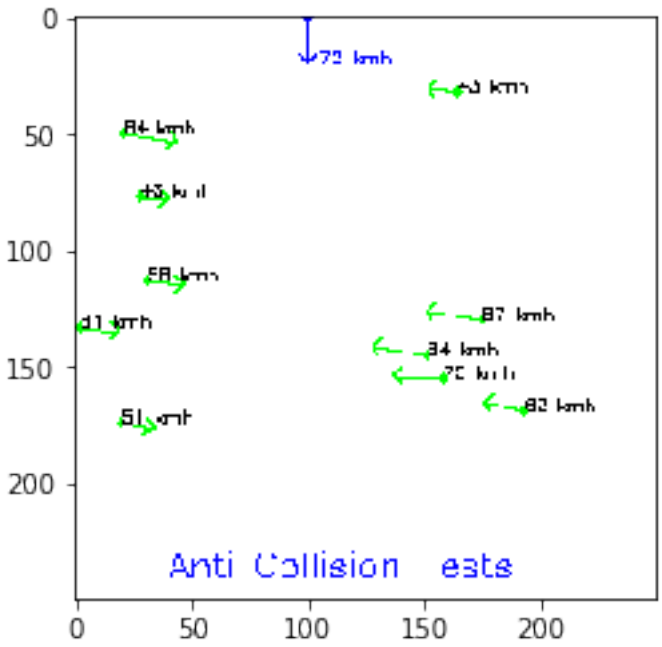
3 DQN network: start experiments with a simple DNN

[5]:
```
class DQN(nn.Module):

    def __init__(self, inputs=44, outputs=5):
        super(DQN, self).__init__()
        self.fc1 = nn.Linear(inputs, 100)
        self.fc2 = nn.Linear(100, 100)
        self.fc3 = nn.Linear(100, outputs)

    def forward(self, x):
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x
```

[6]:
```
# utility to conver numpy arrays to torch arrays
# On GPU if possible, with floats 32 bits
def numpy_to_torch(state):
    s = torch.from_numpy(state).to(device)
    # unsqueeze(0) to add a batch dim
    s = s.unsqueeze(0).to(device).float()
```

```
    return s
```

[7]:
```
BATCH_SIZE = 128
GAMMA = 0.999
EPS_START = 0.9
EPS_END = 0.05
EPS_DECAY = 200
TARGET_UPDATE = 10

# Get number of actions from gym action space
n_actions = env.action_space.n
action = 0
obs, reward, done, info = env.step(action)
env.reset()
img = env.render()
show_img(img)

n_feats = env.observation_space.shape[0] # ego [x,y,vx,vy] + 10 cars [x,y,vx,vy]
assert n_feats == 44 # should be 44 ...

policy_net = DQN(n_feats, n_actions).to(device)
target_net = DQN(n_feats, n_actions).to(device)

policy_net = policy_net.float()
target_net = target_net.float()

target_net.load_state_dict(policy_net.state_dict())
target_net.eval()

optimizer = optim.RMSprop(policy_net.parameters())
memory = ReplayMemory(10000)
steps_done = 0
```



4 Epsilon-greedy exploration

[8]:
```
def select_action(state):
    global steps_done
    sample = random.random()
    eps_threshold = EPS_END + (EPS_START - EPS_END) * \
        math.exp(-1. * steps_done / EPS_DECAY)
    steps_done += 1
    if sample > eps_threshold:
        with torch.no_grad():
            return torch.argmax(policy_net(state)).view(1, 1)
    else:
        return torch.tensor([[random.randrange(n_actions)]], device=device,␣
↪dtype=torch.long)
```

5 Optimize with Huber Loss

[9]:
```
def optimize_model():
    if len(memory) < BATCH_SIZE:
        return
    transitions = memory.sample(BATCH_SIZE)
```

```
    # Transpose the batch (see https://stackoverflow.com/a/19343/3343043 for
    # detailed explanation). This converts batch-array of Transitions
    # to Transition of batch-arrays.
    batch = Transition(*zip(*transitions))

    # Compute a mask of non-final states and concatenate the batch elements
    # (a final state would've been the one after which simulation ended)
    non_final_mask = torch.tensor(tuple(map(lambda s: s is not None, batch.
↪next_state)), device=device, dtype=torch.bool)
    non_final_next_states = torch.cat([s for s in batch.next_state if s is not␣
↪None])
    state_batch = torch.cat(batch.state)
    action_batch = torch.cat(batch.action)
    reward_batch = torch.cat(batch.reward)

    # Compute Q(s_t, a) - the model computes Q(s_t), then we select the
    # columns of actions taken. These are the actions which would've been taken
    # for each batch state according to policy_net
    state_action_values = policy_net(state_batch).gather(1, action_batch)

    # Compute V(s_{t+1}) for all next states.
    # Expected values of actions for non_final_next_states are computed based
    # on the "older" target_net; selecting their best reward with max(1)[0].
    # This is merged based on the mask, such that we'll have either the expected
    # state value or 0 in case the state was final.
    next_state_values = torch.zeros(BATCH_SIZE, device=device)
    next_state_values[non_final_mask] = target_net(non_final_next_states).
↪max(1)[0].detach()
    # Compute the expected Q values
    expected_state_action_values = (next_state_values * GAMMA) + reward_batch

    # Compute Huber loss
    loss = F.smooth_l1_loss(state_action_values, expected_state_action_values.
↪unsqueeze(1))
    print("loss {}".format(loss))

    # Optimize the model
    optimizer.zero_grad()
    loss.backward()
    for param in policy_net.parameters():
        param.grad.data.clamp_(-1, 1)
    optimizer.step()
```

6 Training example (on just a few episodes)

Some usefull references to follow up ideas and code: * Playing Atari with Deep Reinforcement Learning, V. Mnih et al., NIPS Workshop, 2013: https://arxiv.org/abs/1312.5602
* Human-level control through deep reinforcement learning, V. Mnih et al., Nature, 2015: https://deepmind.com/research/publications/human-level-control-through-deep-reinforcement-learning
* Automated Speed and Lane Change Decision Making using Deep Reinforcement Learning, Carl-Johan Hoel et al., ITSC, 2018:
https://arxiv.org/abs/1803.10056
* Pytorch dqn starter code: https://pytorch.org/tutorials/intermediate/reinforcement_q_learning.html

We typically have to train for 10 million steps.
But in this notebook, as an illustration, we just run for around 1000 steps.
We confirm that this problem is much more difficult to solve than eg cartpole.
And is not trivially solved or learned in a few minutes.

[10]:
```
Transition = namedtuple('Transition', ('state', 'action', 'next_state',␣
↪'reward'))

res = 0
num_episodes = 50
for i_episode in range(num_episodes):
    # Initialize the environment and state
    state = env.reset()
    state = numpy_to_torch(state)
    cumulated_reward = 0
    images = []

    for t in count():
        # Select and perform an action
        action = select_action(state)
        next_state, reward, done, info = env.step(action.item())
        next_state = numpy_to_torch(next_state)
        cumulated_reward += reward
        print("Step {}: action={} reward={} done={} info={}".format(t, action,␣
↪reward, done, info))
        img = env.render()
        images.append(img)
        reward = torch.tensor([reward], device=device)

        # Observe new state
        if done:
            #pdb.set_trace()
            next_state = None
            res += cumulated_reward
```

```
            print("End of episode {} with cumulated_reward {}".
 ↪format(i_episode, cumulated_reward))

        # Store the transition in memory
        memory.push(state, action, next_state, reward)

        # Move to the next state
        state = next_state

        # Perform one step of the optimization (on the target network)
        optimize_model()
        if done:
            print('done!')
            #episode_durations.append(t + 1)
            #plot_durations()
            break
    # Update the target network, copying all weights and biases in DQN
    if i_episode % TARGET_UPDATE == 0:
        target_net.load_state_dict(policy_net.state_dict())

print('Completed with an average cumulated reward = {}'.format(res/
 ↪num_episodes))
env.render()
env.close()
```

```
Step 0: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 8: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 9: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 12: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 14: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 16: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 17: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 18: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 19: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 20: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
Step 22: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 23: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 24: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 25: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 26: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 27: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 28: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 29: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 30: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 31: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 32: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 33: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 34: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 35: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 37: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 39: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 40: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 41: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 42: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 43: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 47: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 48: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 49: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 0 with cumulated_reward 950
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 3: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 7: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 9: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 12: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 14: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 16: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 17: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
```

```
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 19: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 22: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 23: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 24: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 25: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 26: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 27: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 28: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 29: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 30: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 31: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 32: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 33: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 34: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 35: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 36: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 37: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 39: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 40: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 41: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 42: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 43: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 44: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 46: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 48: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 49: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 50: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 51: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 52: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
Step 53: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 1 with cumulated_reward 946
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 2: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
Step 10: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 13: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 2 with cumulated_reward -1014
done!
Step 0: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 6: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 29.373708724975586
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 1336.9052734375
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 35.66288757324219
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 28.70933723449707
Step 13: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 43.01341247558594
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 36.19375991821289
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 22.709022521972656
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 22.534870147705078
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 21.10540008544922
Step 18: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 28.924686431884766
Step 19: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 28.98604393005371
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 26.013145446777344
Step 21: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.584007263183594
Step 22: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 19.71440887451172
Step 23: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 28.637948989868164
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 19.73346519470215
Step 25: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
```

```
End of episode 3 with cumulated_reward -1026
loss 33.79412078857422
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 33.951717376708984
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 34.011756896972656
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 33.84203338623047
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 35.45830154418945
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 34.70982360839844
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.23237228393555
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 19.01512336730957
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.665681838989258
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.60625076293945
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 34.64080810546875
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.48247146606445
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.75837326049805
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 33.41853713989258
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 25.823015213012695
Step 14: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 33.571041107177734
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.803253173828125
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 18.074193954467773
Step 17: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 25.252384185791016
Step 18: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 25.725284576416016
Step 19: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.97371292114258
Step 20: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.867717742919922
Step 21: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 25.228702545166016
Step 22: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 33.246482849121094
Step 23: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 25.05978012084961
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 25.442968368530273
Step 25: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 17.16701889038086
Step 26: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.343711853027344
Step 27: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.7120361328125
Step 28: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.290340423583984
Step 29: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.33711624145508
Step 30: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.683996200561523
Step 31: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 1.601675033569336
Step 32: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.01414489746094
Step 33: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.425159454345703
Step 34: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 25.147310256958008
Step 35: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.43345260620117
Step 36: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.339725494384766
Step 37: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.376800537109375
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 25.03415298461914
Step 39: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.39144515991211
Step 40: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.9827880859375
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.734867095947266
Step 42: action=tensor([[0]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 4 with cumulated_reward -1043
loss 32.778079986572266
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.365535736083984
Step 1: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.65065002441406
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 16.621858596801758
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.79711151123047
Step 4: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.475128173828125
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 33.08466339111328
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.09381866455078
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 25.075939178466797
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 25.01520538330078
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.30645751953125
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.825368881225586
Step 11: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 40.33074951171875
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.6965274810791
Step 13: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 17.236169815063477
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.18575668334961
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.842599868774414
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 9.047810554504395
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 9.055529594421387
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.83079719543457
Step 19: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.740055084228516
Step 20: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.97654914855957
Step 21: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 9.257465362548828
Step 22: action=tensor([[3]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 5 with cumulated_reward -1023
loss 16.750307083129883
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 40.270931243896484
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 40.22040557861328
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 40.29167938232422
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.32292938232422
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.70253562927246
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.917816162109375
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.5662841796875
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.983705520629883
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 40.20685577392578
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.4266357421875
Step 10: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.80709457397461
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.433053970336914
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.62929916381836
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.497774124145508
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.51289939880371
Step 15: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.607833862304688
Step 16: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.66645050048828
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.34462356567383
Step 18: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.753782272338867
Step 19: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.13638687133789
Step 20: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 9.038254737854004
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.436735153198242
Step 22: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 9.015275955200195
Step 23: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.821168899536133
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.79762077331543
Step 25: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.04692077636719
Step 26: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 24.4108829498291
Step 27: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.910655975341797
Step 28: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.429744720458984
Step 29: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.974714279174805
Step 30: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.6304988861084
Step 31: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.822007179260254
Step 32: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.296220779418945
Step 33: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.560707092285156
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.43568992614746
Step 35: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.97519874572754
Step 36: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.726091384887695
Step 37: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 40.186336517333984
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.353050231933594
Step 39: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.40748596191406
Step 40: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.729507446289062
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.665094375610352
Step 42: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.13843536376953
Step 43: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.176101684570312
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.406232833862305
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.558218002319336
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.79500961303711
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.36427879333496
Step 48: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 6 with cumulated_reward 951
loss 31.963926315307617
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 39.83409118652344
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.264816284179688
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 34.257896423339844
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.55941390991211
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.132917404174805
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.15509796142578
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.21413803100586
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.693540573120117
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.89398193359375
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.94464111328125
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.603331565856934
Step 11: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.187694549560547
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.916479110717773
Step 13: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.474618911743164
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.283287048339844
Step 15: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.03132629394531
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.28648376464844
Step 17: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.105072021484375
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.05048370361328
Step 19: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.615432739257812
Step 20: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.685081481933594
Step 21: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.446842193603516
Step 22: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.4141845703125
Step 23: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.154605865478516
Step 24: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
```

```
End of episode 7 with cumulated_reward -1025
loss 16.773944854736328
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.91740417480469
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 1.1166507005691528
Step 2: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.782169342041016
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.54692840576172
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.89980697631836
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 40.13678741455078
Step 6: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.37059783935547
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.300020217895508
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.127506256103516
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.984844207763672
Step 10: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.210752487182617
Step 11: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.811765670776367
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.343286514282227
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.620445251464844
Step 14: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.212928771972656
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.888620376586914
Step 16: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.519336700439453
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.83930015563965
Step 18: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.27252197265625
Step 19: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.068580627441406
Step 20: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.315298080444336
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.612586975097656
Step 22: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 24.256023406982422
Step 23: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.673538208007812
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 40.01090621948242
Step 25: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.450218200683594
Step 26: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 40.14035415649414
Step 27: action=tensor([[3]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 8 with cumulated_reward -1028
loss 23.91966438293457
done!
Step 0: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.298320770263672
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.769287109375
Step 2: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.276466369628906
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.908973693847656
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.560319900512695
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.29130744934082
Step 6: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.16851043701172
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.390098571777344
Step 8: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.754615783691406
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.83946418762207
Step 10: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.349504470825195
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.292705535888672
Step 12: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.43895721435547
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.135622024536133
Step 14: action=tensor([[3]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 9 with cumulated_reward -1015
loss 8.677213668823242
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.41313934326172
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.8756046295166
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.541711807250977
Step 3: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.964447021484375
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.761226654052734
Step 5: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.316810607910156
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.282222747802734
Step 7: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.939857482910156
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.984128952026367
Step 9: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.14873504638672
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.02558898925781
Step 11: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.07350158691406
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.69068908691406
Step 13: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.475543975830078
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.194351196289062
Step 15: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.371295928955078
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.72696304321289
Step 17: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.69681167602539
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.447067260742188
Step 19: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.72875213623047
Step 20: action=tensor([[3]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 10 with cumulated_reward -1021
loss 47.42744445800781
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.425100326538086
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.81592559814453
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.91178894042969
Step 3: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 32.03749084472656
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.41336441040039
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.830408096313477
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.90363121032715
Step 7: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.091344833374023
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.495250701904297
Step 9: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.72279167175293
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.545333862304688
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.410531997680664
Step 12: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.29193878173828
Step 13: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.763229370117188
Step 14: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.707773208618164
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.120445251464844
Step 16: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.579933166503906
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.452674865722656
Step 18: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.793926239013672
Step 19: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.904314041137695
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.835329055786133
Step 21: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.732484817504883
Step 22: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.608890533447266
Step 23: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.1827449798584
Step 24: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.47026824951172
Step 25: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.887863159179688
Step 26: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 62.73604202270508
Step 27: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 16.168201446533203
Step 28: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 15.81545352935791
Step 29: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.95101547241211
Step 30: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.051355361938477
Step 31: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.716251373291016
Step 32: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.03709602355957
Step 33: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.149738311767578
Step 34: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 0.5173826217651367
Step 35: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.29570198059082
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.42988967895508
Step 37: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.667030334472656
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 55.071407318115234
Step 39: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.70862579345703
Step 40: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.5902156829834
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.136180877685547
Step 42: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.596599578857422
Step 43: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.669883728027344
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.386253356933594
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.91250228881836
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.467323303222656
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.07172966003418
Step 48: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.929168701171875
Step 49: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.858287811279297
Step 50: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.040605545043945
Step 51: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 32.1346435546875
Step 52: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.815710067749023
Step 53: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.919639587402344
Step 54: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.62578201293945
Step 55: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.1995849609375
Step 56: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.654903411865234
Step 57: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.11192321777344
Step 58: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.59466552734375
Step 59: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.9930305480957
Step 60: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.41360092163086
Step 61: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.434414863586426
Step 62: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.798166275024414
Step 63: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.494998931884766
Step 64: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 55.11817932128906
Step 65: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.984336853027344
Step 66: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 0.5128107666969299
Step 67: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.989341735839844
Step 68: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.89833068847656
Step 69: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 11 with cumulated_reward 930
loss 8.456974983215332
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.256370544433594
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.137883186340332
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.39222717285156
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.244266510009766
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 23.873125076293945
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.527694702148438
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.844186782836914
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.368258476257324
Step 8: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.055362701416016
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.56973648071289
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.4753532409668
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.38323974609375
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.46155548095703
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.660663604736328
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.817577362060547
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.75249671936035
Step 16: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.6854248046875
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 38.99700927734375
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.589906692504883
Step 19: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.725522994995117
Step 20: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.813108444213867
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.39434051513672
Step 22: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.115652084350586
Step 23: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 54.66499328613281
Step 24: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 63.344871520996094
Step 25: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.113483428955078
Step 26: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.663230895996094
Step 27: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.4349365234375
Step 28: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.552040100097656
Step 29: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.78714942932129
Step 30: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.848529815673828
Step 31: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.812015533447266
Step 32: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.543359756469727
Step 33: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.414779663085938
Step 34: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 46.79502487182617
Step 35: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.250465393066406
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 0.5144152045249939
Step 37: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.18545913696289
Step 38: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.081462860107422
Step 39: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.075618743896484
Step 40: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.601329803466797
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.941595077514648
Step 42: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.049907684326172
Step 43: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.022998809814453
Step 44: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.824668884277344
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.938114166259766
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.297706604003906
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.920754432678223
Step 48: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.0380916595459
Step 49: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.503822326660156
Step 50: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.648462295532227
Step 51: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.60104751586914
Step 52: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 39.52106857299805
Step 53: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.916046142578125
Step 54: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.212055206298828
Step 55: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.83679962158203
Step 56: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.155765533447266
Step 57: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.40616798400879
Step 58: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 12 with cumulated_reward 941
loss 31.50596809387207
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.986553192138672
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.547264099121094
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.431476593017578
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.760796546936035
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.884979248046875
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.8260383605957
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.888914108276367
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.19607925415039
Step 8: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.443981170654297
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.766624450683594
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.82671356201172
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.381793975830078
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.07265853881836
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.601774215698242
Step 14: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.143768310546875
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.365970611572266
Step 16: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 46.96907424926758
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.371971130371094
Step 18: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.122825622558594
Step 19: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.124780654907227
Step 20: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.257753372192383
Step 21: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.4340877532959
Step 22: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.758251190185547
Step 23: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.275724411010742
Step 24: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.000568389892578
Step 25: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 38.564476013183594
Step 26: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.82347869873047
Step 27: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 15.668623924255371
Step 28: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.070884704589844
Step 29: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.06035614013672
Step 30: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.59674835205078
Step 31: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.088500022888184
Step 32: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.5794734954834
Step 33: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.49285125732422
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.159515380859375
Step 35: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.827852249145508
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.81043243408203
Step 37: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.09488868713379
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.140684127807617
Step 39: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.869224548339844
Step 40: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 39.51449203491211
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.401460647583008
Step 42: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.5892333984375
Step 43: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.33584976196289
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.904112815856934
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.95991516113281
Step 46: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.936408996582031
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.31501579284668
Step 48: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.98448944091797
Step 49: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.41685676574707
Step 50: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.534170150756836
Step 51: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 13 with cumulated_reward 948
loss 23.18862533569336
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 38.990482330322266
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 62.37038803100586
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 46.68739318847656
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.241981506347656
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 0.6404271721839905
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.398468017578125
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.641584396362305
Step 7: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.07900619506836
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.5268611907959
Step 9: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.30206298828125
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.001909255981445
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 54.456329345703125
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.807869911193848
Step 13: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.878955841064453
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 38.579139709472656
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.553342819213867
Step 16: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 54.16596984863281
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 54.37342071533203
Step 18: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 30.584047317504883
Step 19: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.688854217529297
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.289630889892578
Step 21: action=tensor([[0]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 14 with cumulated_reward -1022
loss 31.382686614990234
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 38.35809326171875
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 22.68295669555664
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 30.096298217773438
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.677127838134766
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.7979736328125
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.06061363220215
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.80568504333496
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.269119262695312
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 38.48575973510742
Step 9: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.468875885009766
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.733816146850586
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.903169631958008
Step 12: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.770069122314453
Step 13: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.84143829345703
Step 14: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.61898136138916
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.15707015991211
Step 16: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.68621826171875
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 30.53548240661621
Step 18: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.478713989257812
Step 19: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.892717361450195
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.58509063720703
Step 21: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.4814453125
Step 22: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.62877655029297
Step 23: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.745136260986328
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.788423538208008
Step 25: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.263203620910645
Step 26: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.79419708251953
Step 27: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 30.21923065185547
Step 28: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.75298309326172
Step 29: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.286975860595703
Step 30: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.5126838684082
Step 31: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 46.831451416015625
Step 32: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 22.580780029296875
Step 33: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.188899993896484
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.215490341186523
Step 35: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 21.97701644897461
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.97949981689453
Step 37: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.523962020874023
Step 38: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 30.031461715698242
Step 39: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.372920989990234
Step 40: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.88108253479004
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.234546661376953
Step 42: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.890514373779297
Step 43: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.1309928894043
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.959592819213867
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.67099380493164
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 37.64911651611328
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.31210708618164
Step 48: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 46.84463882446289
Step 49: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.587268829345703
Step 50: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.246621131896973
Step 51: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.527923583984375
Step 52: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 29.601444244384766
Step 53: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 13.154345512390137
Step 54: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.58197021484375
Step 55: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.910603523254395
Step 56: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.563093185424805
Step 57: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.44769102334976196
Step 58: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 15 with cumulated_reward 941
loss 16.70688247680664
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 21.0860652923584
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.18290901184082
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.65111541748047
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.498723030090332
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.16344451904297
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 13.75059700012207
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 20.520736694335938
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 28.802993774414062
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.430883407592773
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 0.6676558256149292
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 20.421817779541016
Step 11: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.21755027770996
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.956748962402344
Step 13: action=tensor([[3]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 16 with cumulated_reward -1014
loss 16.50264549255371
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.921299934387207
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.191213607788086
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 36.44929504394531
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.486221313476562
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.75664710998535
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.087825775146484
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 20.847763061523438
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.57599639892578
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.49165403842926025
Step 9: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 38.997013092041016
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.334136962890625
Step 11: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.048206329345703
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 12.524223327636719
Step 13: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 27.700754165649414
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.524479866027832
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 35.65169906616211
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.686077117919922
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 27.69955825805664
Step 18: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.218412399291992
Step 19: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 50.4929084777832
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.86322021484375
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.119842529296875
Step 22: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.553102493286133
Step 23: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.774751663208008
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.364908218383789
Step 25: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 12.123244285583496
Step 26: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.270950317382812
Step 27: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.331592559814453
Step 28: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 11.518704414367676
Step 29: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.08465576171875
Step 30: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.82212448120117
Step 31: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 41.792823791503906
Step 32: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.35019874572754
Step 33: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
```

```
End of episode 17 with cumulated_reward -1034
loss 39.34343338012695
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.7675895690918
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 49.422279357910156
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.430198669433594
Step 3: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 33.677345275878906
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 17.02764320373535
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 49.217838287353516
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.6226478815078735
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.340229034423828
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.35645294189453
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.948328018188477
Step 10: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 18 with cumulated_reward -1011
loss 31.916423797607422
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.394240379333496
Step 1: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.119417190551758
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.102577209472656
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.1397705078125
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.150432586669922
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.83444595336914
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.193817138671875
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.188032150268555
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 19.2640380859375
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.735504150390625
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.752880096435547
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.321537017822266
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 10.801475524902344
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.314152717590332
Step 14: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.831741333007812
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.93935012817383
Step 16: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.282167434692383
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.777510643005371
Step 18: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.78589630126953
Step 19: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 15.999656677246094
Step 20: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.79140853881836
Step 21: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.952022552490234
Step 22: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.334156036376953
Step 23: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 26.25775146484375
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.047168731689453
Step 25: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.391529083251953
Step 26: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.00082778930664
Step 27: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.2735710144043
Step 28: action=tensor([[0]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 19 with cumulated_reward -1029
loss 31.611982345581055
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.835115432739258
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.42970085144043
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.048267364501953
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 41.77134704589844
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 15.9025239944458
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 18.227130889892578
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.77782440185547
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.885041236877441
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.326309204101562
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 7.994740009307861
Step 10: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 20 with cumulated_reward -1011
loss 56.07024002075195
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.53669738769531
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 25.083236694335938
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 62.231346130371094
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.713634490966797
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 40.549095153808594
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.986454010009766
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.43150329589844
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.452091217041016
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.71844482421875
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.312259674072266
Step 10: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.803455352783203
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 54.96700668334961
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 17.612075805664062
Step 13: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 21 with cumulated_reward -1014
loss 47.12397766113281
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.667999267578125
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 17.436891555786133
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.771347045898438
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 46.89943313598633
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.717065811157227
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.96995735168457
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.49711799621582
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.34440231323242
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.930723190307617
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.522323608398438
Step 10: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.462503433227539
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.46540069580078
Step 12: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.731897354125977
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.74256134033203
Step 14: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.052074432373047
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.788211822509766
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.404563903808594
Step 17: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 55.3702278137207
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.6816397309303284
Step 19: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.0244083404541
Step 20: action=tensor([[3]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 22 with cumulated_reward -1021
loss 16.03181266784668
done!
Step 0: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.45526123046875
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.252723693847656
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.756500244140625
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 24.019765853881836
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.62439727783203
Step 5: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.96640396118164
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.02355194091797
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.0005989074707
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.876869201660156
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.6176931858062744
Step 10: action=tensor([[0]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 23 with cumulated_reward -1011
loss 23.817075729370117
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.848209381103516
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.903995513916016
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 38.93610382080078
Step 3: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.42518424987793
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.544204711914062
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.13749313354492
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.743778228759766
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.144981384277344
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.489253997802734
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.06393051147461
Step 10: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 24 with cumulated_reward -1011
loss 23.72132682800293
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 41.54421615600586
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.321340560913086
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 9.389485359191895
Step 3: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 63.35859680175781
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.38823699951172
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.450611114501953
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 55.927146911621094
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 17.261987686157227
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 55.250064849853516
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.81553840637207
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.536487579345703
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.583003997802734
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.52550220489502
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.057920455932617
Step 14: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.372369766235352
Step 15: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.259883880615234
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.18369483947754
Step 17: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 25 with cumulated_reward -1018
loss 23.582969665527344
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.22764778137207
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.449195861816406
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.69263458251953
Step 3: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.532352447509766
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.980918884277344
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.768892288208008
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.339195251464844
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.156414031982422
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 24.059555053710938
Step 9: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.301284790039062
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.94989776611328
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.875603675842285
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.86018180847168
Step 13: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.464189529418945
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.087825775146484
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.767999649047852
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.18624496459961
Step 17: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.47107219696045
Step 18: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.852636337280273
Step 19: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.95208740234375
Step 20: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.797542572021484
Step 21: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.602008819580078
Step 22: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.882373809814453
Step 23: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.906105041503906
Step 24: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.473499298095703
Step 25: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.837181091308594
Step 26: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.505382537841797
Step 27: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.670398712158203
Step 28: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.03802490234375
Step 29: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.971717834472656
Step 30: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.820009231567383
Step 31: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.958521842956543
Step 32: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 62.573970794677734
Step 33: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.688735961914062
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.67025375366211
Step 35: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.73685836791992
Step 36: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.36652374267578
Step 37: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.770292282104492
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.771387100219727
Step 39: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.55026626586914
Step 40: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 63.053131103515625
Step 41: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.76556968688965
Step 42: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.338998794555664
Step 43: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.845149993896484
Step 44: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.42438507080078
Step 45: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.35732650756836
Step 46: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.628881454467773
Step 47: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.078845977783203
Step 48: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.77898597717285
Step 49: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.181132316589355
Step 50: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.037403106689453
Step 51: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 56.31418991088867
Step 52: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.90542984008789
Step 53: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.04819869995117
Step 54: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.90445327758789
Step 55: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.696500778198242
Step 56: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.957916259765625
Step 57: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.620899200439453
Step 58: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 55.2735595703125
Step 59: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.4102897644043
Step 60: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.646852493286133
Step 61: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.5365804433822632
Step 62: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.79878807067871
Step 63: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.93617820739746
Step 64: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.11769104003906
Step 65: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.26475524902344
Step 66: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.693084716796875
Step 67: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 55.136287689208984
Step 68: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.079547882080078
Step 69: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.562580108642578
Step 70: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.4587459564209
Step 71: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.420166015625
Step 72: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.89884567260742
Step 73: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.651256561279297
Step 74: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.04652786254883
Step 75: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.884187698364258
Step 76: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.11884689331055
Step 77: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.278873443603516
Step 78: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.82673454284668
Step 79: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.47649383544922
Step 80: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.853851318359375
Step 81: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.54983901977539
Step 82: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.836572647094727
Step 83: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.44917297363281
Step 84: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.599037170410156
Step 85: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 55.272525787353516
Step 86: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.49126434326172
Step 87: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.70613670349121
Step 88: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.265941619873047
Step 89: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.135174751281738
Step 90: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.329681396484375
Step 91: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.203983306884766
Step 92: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.09812927246094
Step 93: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.50902557373047
Step 94: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.347375869750977
Step 95: action=tensor([[0]], device='cuda:0') reward=999 done=True info=success
End of episode 26 with cumulated_reward 904
loss 38.9100456237793
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 20.431032180786133
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.94293975830078
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.318660736083984
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.986759185791016
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.8094596862793
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.52953338623047
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.818880081176758
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 70.18760681152344
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.636455535888672
Step 9: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 27 with cumulated_reward -1010
loss 32.35502243041992
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.91100311279297
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.335410118103027
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 55.204498291015625
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.0451774597168
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.357819557189941
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.5485954284668
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.15018653869629
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 62.102752685546875
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.489973068237305
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.014158248901367
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.60323715209961
Step 11: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.73057556152344
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.87604522705078
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.857418060302734
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.154512405395508
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.54151153564453
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.70162582397461
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.264394760131836
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.31899642944336
Step 19: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 62.2867546081543
Step 20: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 8.362621307373047
Step 21: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.186065673828125
Step 22: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 55.018165588378906
Step 23: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.101139068603516
Step 24: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.19298553466797
Step 25: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.4304256439209
Step 26: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 46.98594665527344
Step 27: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.703542709350586
Step 28: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.88626480102539
Step 29: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.97709655761719
Step 30: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 30.940326690673828
Step 31: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.989622116088867
Step 32: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.412300109863281
Step 33: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 33.05982971191406
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 30.930614471435547
Step 35: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 62.4807014465332
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.21211814880371
Step 37: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.867609024047852
Step 38: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.896984100341797
Step 39: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 54.94851303100586
Step 40: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.719968795776367
Step 41: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.807498931884766
Step 42: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.50578308105469
Step 43: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.785614013671875
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 16.169084548950195
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.99309730529785
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.56419372558594
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.444530487060547
Step 48: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.32447052001953
Step 49: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.797922134399414
Step 50: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 1.6040167808532715
Step 51: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.321897506713867
Step 52: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.8714599609375
Step 53: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.944469451904297
Step 54: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 15.888508796691895
Step 55: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.630157470703125
Step 56: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 28 with cumulated_reward 943
loss 42.892818450927734
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.270305633544922
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.038644790649414
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 46.40267562866211
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.46290397644043
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.45766830444336
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.486656188964844
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 55.85896682739258
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.968177795410156
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.63580322265625
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.070497512817383
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 15.742603302001953
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.63487243652344
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.92374038696289
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.198617935180664
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.897351861000061
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 33.64710998535156
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.89788055419922
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 9.10101318359375
Step 18: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 15.925521850585938
Step 19: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.777095794677734
Step 20: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.262975692749023
Step 21: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.25111770629883
Step 22: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.17586898803711
Step 23: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.989013671875
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.585933685302734
Step 25: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.735828399658203
Step 26: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.926727294921875
Step 27: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.828207969665527
Step 28: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.08902359008789
Step 29: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.4190616607666
Step 30: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.432100296020508
Step 31: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.34539222717285
Step 32: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.864662170410156
Step 33: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.660053253173828
Step 34: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 8.354409217834473
Step 35: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.956206321716309
Step 36: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 25.79319190979004
Step 37: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.419864654541016
Step 38: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.28394317626953
Step 39: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.675588607788086
Step 40: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.259057998657227
Step 41: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.704265594482422
Step 42: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.001588821411133
Step 43: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.376665115356445
Step 44: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.90211868286133
Step 45: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.444595336914062
Step 46: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.838590621948242
Step 47: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.49744415283203
Step 48: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.440204620361328
Step 49: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 40.50716781616211
Step 50: action=tensor([[3]], device='cuda:0') reward=999 done=True info=success
End of episode 29 with cumulated_reward 949
loss 16.77618980407715
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.619964599609375
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 16.679100036621094
Step 2: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.513479232788086
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.991881370544434
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 12.766368865966797
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.985397338867188
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 23.106449127197266
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.88328742980957
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.871784210205078
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.546539306640625
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.914125442504883
Step 11: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 18.485828399658203
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 33.674102783203125
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.861661911010742
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.212461471557617
Step 15: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.717018127441406
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.631038665771484
Step 17: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 49.20809555053711
Step 18: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.73727035522461
Step 19: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.860170364379883
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.21178436279297
Step 21: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.29922103881836
Step 22: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.677011489868164
Step 23: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 0.9621655344963074
Step 24: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 32.153099060058594
Step 25: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.030216217041016
Step 26: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 47.57297134399414
Step 27: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 40.084259033203125
Step 28: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.22369384765625
Step 29: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 40.1031494140625
Step 30: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 24.31200408935547
Step 31: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.397991180419922
Step 32: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.378116607666016
Step 33: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.017242431640625
Step 34: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.63163757324219
Step 35: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.67167663574219
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.654123306274414
Step 37: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 78.81263732910156
Step 38: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.492645263671875
Step 39: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 24.26090431213379
Step 40: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.848971366882324
Step 41: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.638530731201172
Step 42: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 46.73672103881836
Step 43: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.442313194274902
Step 44: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.819705963134766
Step 45: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 54.854393005371094
Step 46: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.781198501586914
Step 47: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.216991424560547
Step 48: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.383949279785156
Step 49: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.31269836425781
Step 50: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 17.19028663635254
Step 51: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.63137435913086
Step 52: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.766536712646484
Step 53: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 15.86355209350586
Step 54: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 17.371196746826172
Step 55: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.35091209411621
Step 56: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.737932205200195
Step 57: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.28598403930664
Step 58: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.802701950073242
Step 59: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.90391540527344
Step 60: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.384765625
Step 61: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.491493225097656
Step 62: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 17.54195785522461
Step 63: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.846168518066406
Step 64: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.337608337402344
Step 65: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.485687255859375
Step 66: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.459186553955078
Step 67: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.60054016113281
Step 68: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.26496505737305
Step 69: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.783729553222656
Step 70: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 30 with cumulated_reward 929
loss 0.6789461374282837
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.580718994140625
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 30.884004592895508
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 38.627166748046875
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 40.89710998535156
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.045166015625
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 59.88292694091797
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 17.585702896118164
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 10.166505813598633
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.34368896484375
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.67078399658203
Step 10: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 40.586700439453125
Step 11: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 47.64301300048828
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 17.535058975219727
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.580244064331055
Step 14: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 31 with cumulated_reward -1015
loss 32.03644561767578
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.79018020629883
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.35528564453125
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.71829605102539
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.03883934020996
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.47734832763672
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.68317413330078
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 9.528274536132812
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.0311279296875
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 38.68416976928711
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.682478904724121
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.711021423339844
Step 11: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.03685760498047
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 25.193862915039062
Step 13: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.419363021850586
Step 14: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.803525924682617
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.08251190185547
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.772886276245117
Step 17: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.322952270507812
Step 18: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 32 with cumulated_reward -1019
loss 16.35378074645996
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.638075828552246
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.431589126586914
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.67676830291748
Step 3: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.402681350708008
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.777758598327637
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.13791275024414
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.287739753723145
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 54.471519470214844
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.661542892456055
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 38.99052429199219
Step 10: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 1.2820723056793213
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.808990478515625
Step 12: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 33 with cumulated_reward -1013
loss 32.32106399536133
done!
Step 0: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.1124153137207
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.474979400634766
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.88167953491211
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.234609603881836
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 39.02669906616211
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.07075500488281
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.824710845947266
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.797086715698242
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.88136291503906
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.496402740478516
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.59821319580078
Step 11: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.00746536254883
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.59768295288086
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 48.129356384277344
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.41877746582031
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 56.212730407714844
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.327392578125
Step 17: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.007516860961914
Step 18: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 34 with cumulated_reward -1019
loss 31.93485450744629
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.881744384765625
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.417335510253906
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.914142608642578
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 19.16499900817871
Step 4: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 40.02125549316406
Step 5: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 15.843799591064453
Step 6: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.421884536743164
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.432138442993164
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 55.070335388183594
Step 9: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.95477867126465
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.78545379638672
Step 11: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 35 with cumulated_reward -1012
loss 31.42518424987793
done!
Step 0: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.835933685302734
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.67821502685547
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.734503746032715
Step 3: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.22211456298828
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.577285766601562
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.856876373291016
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.497352600097656
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 7.9561920166015625
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.20100975036621
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.996280670166016
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.364952087402344
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.991905212402344
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.3544921875
Step 13: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.753900527954102
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.012229919433594
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.481834411621094
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.793224334716797
Step 17: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 36 with cumulated_reward -1018
loss 16.127599716186523
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 16.133962631225586
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.0665340423584
Step 2: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 70.48937225341797
Step 3: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.492786407470703
Step 4: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.63642120361328
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.765762329101562
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.59553527832031
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.51803207397461
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.111974716186523
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.25251007080078
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.71930503845215
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 25.096134185791016
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 0.8136351704597473
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.959383010864258
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 32.1571044921875
Step 15: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 46.62485122680664
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.71898651123047
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.70363426208496
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.45675277709961
Step 19: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.38174819946289
Step 20: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.368061065673828
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.69795608520508
Step 22: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.546329498291016
Step 23: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.497560501098633
Step 24: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 47.109596252441406
Step 25: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.334321975708008
Step 26: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 63.39780807495117
Step 27: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.4267463684082
Step 28: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 37 with cumulated_reward -1029
loss 32.80785369873047
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.849796295166016
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.03020477294922
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.547500610351562
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.64704132080078
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.045719146728516
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.276500701904297
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.89984130859375
Step 7: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.62102508544922
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.573699951171875
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.835285186767578
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 38.8032112121582
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 30.809425354003906
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 38.91230010986328
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.981915473937988
Step 14: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.652923583984375
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.252784729003906
Step 16: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.64432716369629
Step 17: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.02117156982422
Step 18: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 47.466407775878906
Step 19: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.85254669189453
Step 20: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.903839111328125
Step 21: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.882798194885254
Step 22: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.361228942871094
Step 23: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.27382278442383
Step 24: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 30.9827823638916
Step 25: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.973514556884766
Step 26: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 56.06982421875
Step 27: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.73757553100586
Step 28: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.860118865966797
Step 29: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.860153198242188
Step 30: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.311546325683594
Step 31: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.566696166992188
Step 32: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.90530014038086
Step 33: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.939441680908203
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.27483081817627
Step 35: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.981544494628906
Step 36: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.92093276977539
Step 37: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.85296630859375
Step 38: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.360450744628906
Step 39: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.62027359008789
Step 40: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.647621154785156
Step 41: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.23387908935547
Step 42: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.533710479736328
Step 43: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.432798385620117
Step 44: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.749454498291016
Step 45: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.490501403808594
Step 46: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 17.677640914916992
Step 47: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.58155632019043
Step 48: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.010339736938477
Step 49: action=tensor([[2]], device='cuda:0') reward=999 done=True info=success
End of episode 38 with cumulated_reward 950
loss 46.959312438964844
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.756543159484863
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 38.61189651489258
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 46.87331771850586
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.58756446838379
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.15605545043945
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.02152633666992
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.420310974121094
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.744800567626953
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.56830596923828
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.220245361328125
Step 10: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 39 with cumulated_reward -1011
loss 15.719714164733887
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.182645797729492
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.313151359558105
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.15070724487305
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 39.912254333496094
Step 4: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.36037826538086
Step 5: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.230579376220703
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.23536682128906
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 47.23543167114258
Step 8: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.09295654296875
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.23857307434082
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.363929748535156
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.095983505249023
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.4395637512207
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.77897834777832
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.004318237304688
Step 15: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.913782119750977
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.214786529541016
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.619693756103516
Step 18: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.21816635131836
Step 19: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.95065689086914
Step 20: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.07085609436035
Step 21: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.712059020996094
Step 22: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 40 with cumulated_reward -1023
loss 16.073640823364258
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.405216217041016
Step 1: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.828868865966797
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 63.25403594970703
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 39.98295593261719
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.247394561767578
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.775678634643555
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 61.92405700683594
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.479862213134766
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.553060531616211
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.53338050842285
Step 10: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.26698112487793
Step 11: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 41 with cumulated_reward -1012
loss 8.063267707824707
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.783370971679688
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 47.0386962890625
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.348735809326172
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.270233154296875
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.596113204956055
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 46.76997375488281
Step 6: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.06515884399414
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.85630798339844
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.111892700195312
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.769393920898438
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.00547218322754
Step 11: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.18584442138672
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.979066848754883
Step 13: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.426876068115234
Step 14: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
```

```
End of episode 42 with cumulated_reward -1015
loss 39.28678894042969
done!
Step 0: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.10361862182617
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.34510040283203
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.315839767456055
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 38.99224853515625
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 9.185057640075684
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 8.506897926330566
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 56.07554244995117
Step 7: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.23318099975586
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 55.4161376953125
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.315311431884766
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 61.88471984863281
Step 11: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.015457153320312
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.223541259765625
Step 13: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 70.10533142089844
Step 14: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 43 with cumulated_reward -1015
loss 16.074033737182617
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.960193634033203
Step 1: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.3507080078125
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.78145980834961
Step 3: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.681026458740234
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 23.749130249023438
Step 5: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.54762840270996
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 38.864341735839844
Step 7: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 24.247814178466797
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 62.27503967285156
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.93741226196289
Step 10: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 15.849729537963867
Step 11: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 24.7866153717041
Step 12: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.05017852783203
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.174264907836914
Step 14: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.18968963623047
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 70.17101287841797
Step 16: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.235742568969727
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.183223724365234
Step 18: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 32.1809196472168
Step 19: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.454954147338867
Step 20: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.353452682495117
Step 21: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 38.560760498046875
Step 22: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.559810638427734
Step 23: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.33671188354492
Step 24: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.72577667236328
Step 25: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 46.971343994140625
Step 26: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.27310562133789
Step 27: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 39.19567108154297
Step 28: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.696361541748047
Step 29: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.211326599121094
Step 30: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.659679412841797
Step 31: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.880386352539062
Step 32: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.44701385498047
Step 33: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.794755935668945
Step 34: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.59640884399414
Step 35: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 0.9310188889503479
Step 36: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 32.25779724121094
Step 37: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.463401794433594
Step 38: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 40.192169189453125
Step 39: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 47.33013153076172
Step 40: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 56.063297271728516
Step 41: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.420581817626953
Step 42: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 54.88865661621094
Step 43: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.40897750854492
Step 44: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 40.42892074584961
Step 45: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.02742385864258
Step 46: action=tensor([[3]], device='cuda:0') reward=999 done=True info=success
End of episode 44 with cumulated_reward 953
loss 32.79957962036133
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 54.68302917480469
Step 1: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 47.77557373046875
Step 2: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.15273666381836
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 47.940673828125
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.230751991271973
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.737138748168945
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 54.47024154663086
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.967103958129883
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 40.80870056152344
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.456205368041992
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 54.359596252441406
Step 11: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 15.747391700744629
Step 12: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.526811599731445
Step 13: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.964385986328125
Step 14: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.576282501220703
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 62.57993698120117
Step 16: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.00236892700195
Step 17: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.906909942626953
Step 18: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.446969985961914
Step 19: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.86278533935547
Step 20: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.495765686035156
Step 21: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.48869323730469
Step 22: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.534486770629883
Step 23: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 54.49307632446289
Step 24: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.74437141418457
Step 25: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.84223747253418
Step 26: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.087127685546875
Step 27: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.199214935302734
Step 28: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.797510147094727
Step 29: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 8.458169937133789
Step 30: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 47.27336120605469
Step 31: action=tensor([[2]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 45 with cumulated_reward -1032
loss 47.45534133911133
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 54.2259521484375
Step 1: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.24073028564453
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 23.8339900970459
Step 3: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.724979400634766
Step 4: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 47.910091400146484
Step 5: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 16.315275192260742
Step 6: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 39.16242218017578
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.020404815673828
Step 8: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 24.019676208496094
Step 9: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.33572006225586
Step 10: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 46 with cumulated_reward -1011
loss 15.577473640441895
done!
Step 0: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 69.53410339355469
Step 1: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 31.92112922668457
Step 2: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.77811050415039
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.86431121826172
Step 4: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 8.588321685791016
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 38.452598571777344
Step 6: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 16.378164291381836
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.795037269592285
Step 8: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 31.665006637573242
Step 9: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
```

```
loss 31.584123611450195
Step 10: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 23.784446716308594
Step 11: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.16636276245117
Step 12: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 32.93238067626953
Step 13: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 47.113197326660156
Step 14: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 54.50726318359375
Step 15: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 47 with cumulated_reward -1016
loss 15.559542655944824
done!
Step 0: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 9.589560508728027
Step 1: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 39.31410217285156
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.32167053222656
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.264463424682617
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 8.158730506896973
Step 5: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.888803482055664
Step 6: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 45.9075927734375
Step 7: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 15.053753852844238
Step 8: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 15.800653457641602
Step 9: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 7.963082790374756
Step 10: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.168968200683594
Step 11: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 55.29717254638672
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 16.077983856201172
Step 13: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 24.105422973632812
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 15.877126693725586
Step 15: action=tensor([[4]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 48 with cumulated_reward -1016
loss 8.213040351867676
```

```
done!
Step 0: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.560218811035156
Step 1: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.608455657958984
Step 2: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 30.812517166137695
Step 3: action=tensor([[4]], device='cuda:0') reward=-1 done=False info={}
loss 31.89175796508789
Step 4: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 39.49263000488281
Step 5: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 55.69327926635742
Step 6: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 39.56972122192383
Step 7: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.06266975402832
Step 8: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 31.79653549194336
Step 9: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 8.592671394348145
Step 10: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 23.5947322845459
Step 11: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 38.400428771972656
Step 12: action=tensor([[0]], device='cuda:0') reward=-1 done=False info={}
loss 46.21432113647461
Step 13: action=tensor([[3]], device='cuda:0') reward=-1 done=False info={}
loss 23.25432014465332
Step 14: action=tensor([[2]], device='cuda:0') reward=-1 done=False info={}
loss 16.83587646484375
Step 15: action=tensor([[1]], device='cuda:0') reward=-1 done=False info={}
loss 31.65509033203125
Step 16: action=tensor([[1]], device='cuda:0') reward=-1001 done=True info=fail
End of episode 49 with cumulated_reward -1017
loss 17.51702880859375
done!
Completed with an average cumulated reward = -509.18
```

With just a few episodes of training, the trajectory is not collision free most of time.

For a failure (collision) we get a reward of -1000 and end an episode.
For a succes (target reached) we get a reward of 999.
Average return after 100 episodes training: -509.18

[ ]: