**实验配置**

第一步： 微调 PEL

第二步： 微调 ICD

Step 1.收集数据：

以target=80/70，收集200/400轮数据（switch=50）（与z构成连接）

（用finetune后的PEL和pretrained的ICD与环境交互四次）

Step2.训练阶段：

用这些数据和离线数据一起微调ICD，

数据处理（高回报高比例训练）/（均匀训练）

对数据处理和离线数据一样/用在线数据处理

Step3.测试阶段

Target=40/80/181

|  |  |  |
| --- | --- | --- |
|  | Target=70收集数据 | Target=80收集数据 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Target =80 收集200轮数据 均匀训练，离线数据处理

model/PA-unseen-3oppo-ours-a1-l1-W5-K20-20240626105103

Target=70 收集400轮数据 均匀训练，离线数据处理

model/PA-unseen-3oppo-ours-a1-l1-W5-K20-20240627005443

|  |  |  |
| --- | --- | --- |
| Target = 35 | Target=70 | Target=181 |
| 34.97 | 35.30 | 31.53 |

Target=70 收集400轮数据 高回报高比例训练，离线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627022005

|  |  |  |
| --- | --- | --- |
| Target = 35 | Target = 70 | Target =181 |
| 35.192 | 35.337 | 34.71 |

Target=80 收集400轮数据 高回报高比例训练，离线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627022120

|  |  |  |
| --- | --- | --- |
| Target = 40 | Target= 80 | Target=181 |
| 34.921 | 35.172 | 34.162 |

Target=80 收集400轮数据 均匀训练， 离线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627022841

|  |  |  |
| --- | --- | --- |
| Target = 40 | Target = 80 | Target = 181 |
| 35.1 | 35.406 | 33.745 |

Target=70 均匀训练 在线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627024453

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 30.35 | 30.034 | 13.877 |

Target=80 均匀训练 在线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627024628

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 30.174 | 29.855 | 23.620 |

Target=80 高回报高比例训练，在线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627024929

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.857 | 29.130 | 23.89 |

Target=70 高回报高比例训练，在线数据处理

model/PA-6oppo-ours-a1-l1-W5-K20-20240627025114

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.374 | 29.511 | 24.569 |

Target=80 收集数据， 从头训练一个 PEL,ICD 均匀训练

model/PA-6oppo-ours-a1-l1-W5-K20-20240627030416

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.813 | 29.737 | 29.209 |

Target=80 收集数据，从头训练一个 PEL ,ICD 高回报高比例训练

model/PA-6oppo-ours-a1-l1-W5-K20-20240627030958

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 30.131 | 30.410 | 30.38 |

Target=80 收集数据，训练一个ICD，以 PEL fine tune 均匀训练

model/PA-6oppo-ours-a1-l1-W5-K20-20240627044851

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.949 | 29.361 | 28.395 |

Target=80 收集数据，训练一个ICD 高回报高比例训练

model/PA-6oppo-ours-a1-l1-W5-K20-20240627045059

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 30.196 | 30.520 | 30.681 |

第二轮：

再次训练

PA-6oppo-ours-a1-l1-W5-K20-20240627113506

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.538 | 29.215 | 28.913 |

第一轮：

纯用好的训练

PA-6oppo-ours-a1-l1-W5-K20-20240627114603

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.297 | 29.896 | 30.82 |

比例5：1

PA-6oppo-ours-a1-l1-W5-K20-20240627114844

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 29.454 | 29.492 | 29.367 |

比例3：1

PA-6oppo-ours-a1-l1-W5-K20-20240627114953

比例2：1

PA-6oppo-ours-a1-l1-W5-K20-20240627115118

收集

收集 target=70 400轮数据

collected\_data\_target70\_400.pkl

超过35的比例有 378：422

收集 target=80 400轮数据

collected\_data\_target80\_400.pkl

超过40的比例有 354：446