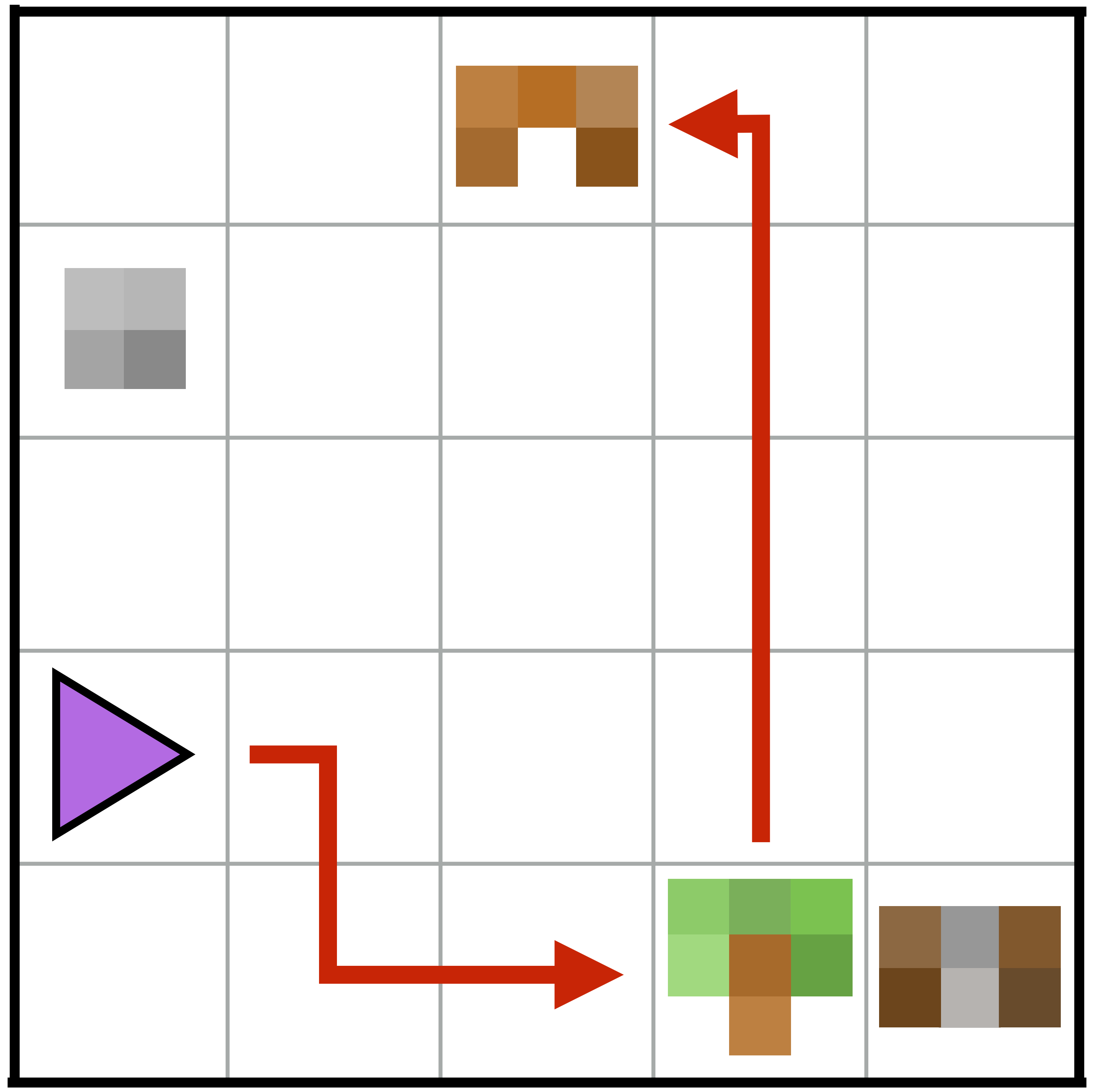
Modular multitask reinforcement learning with policy sketches

Berkeley
NLP

Jacob Andreas, Sergey Levine and Dan Klein

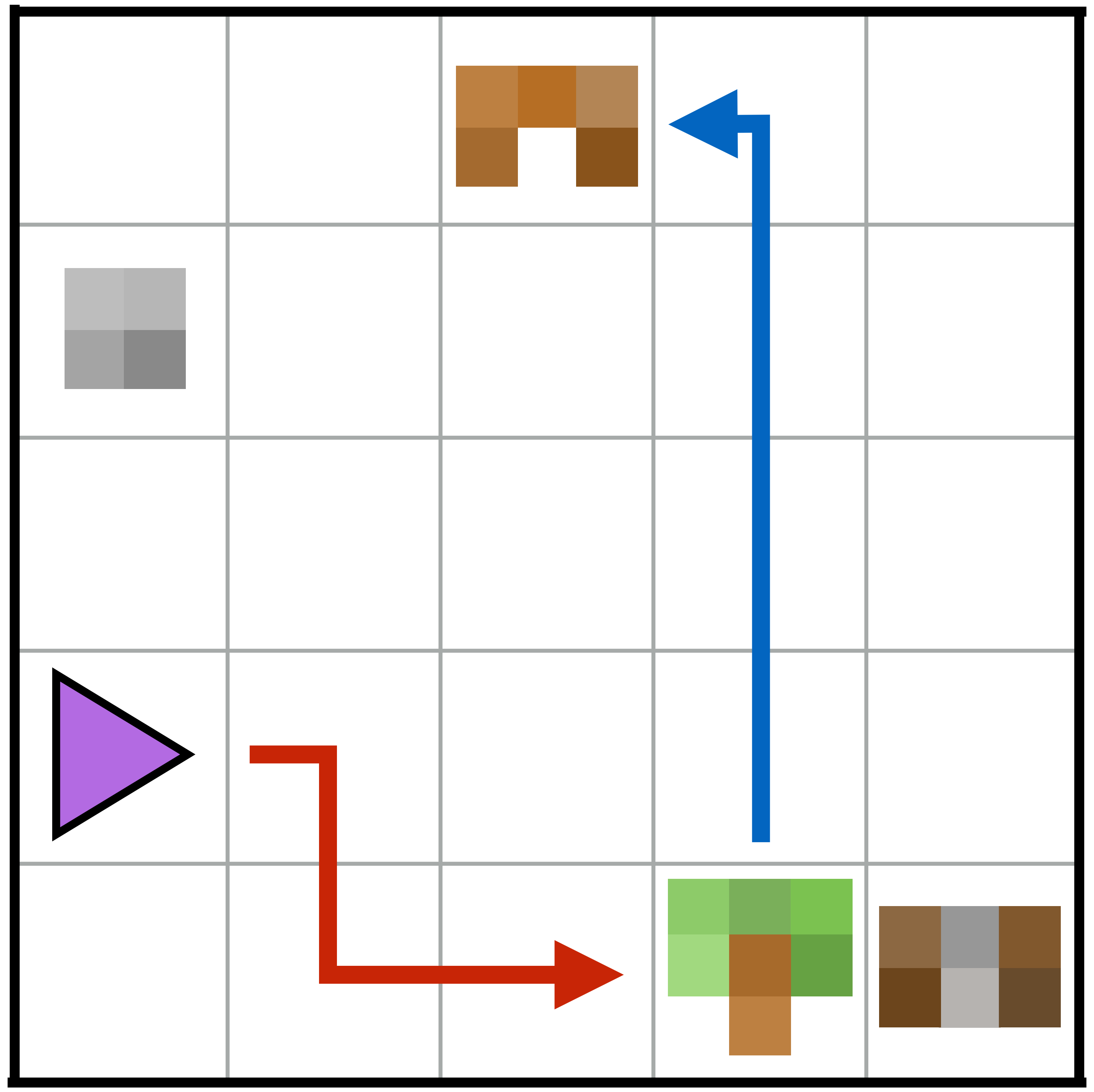
The learning problem

`make planks`

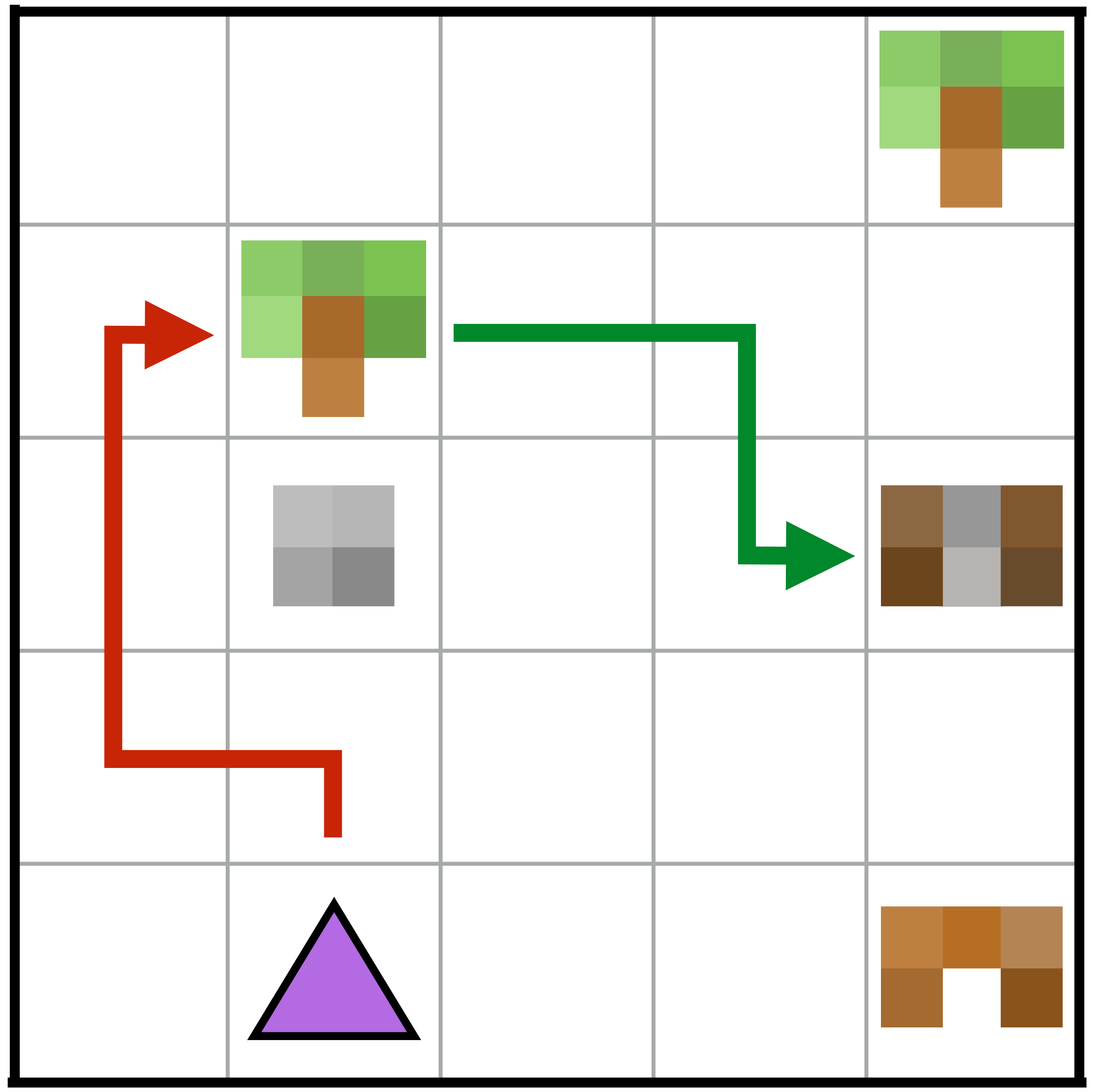


The learning problem

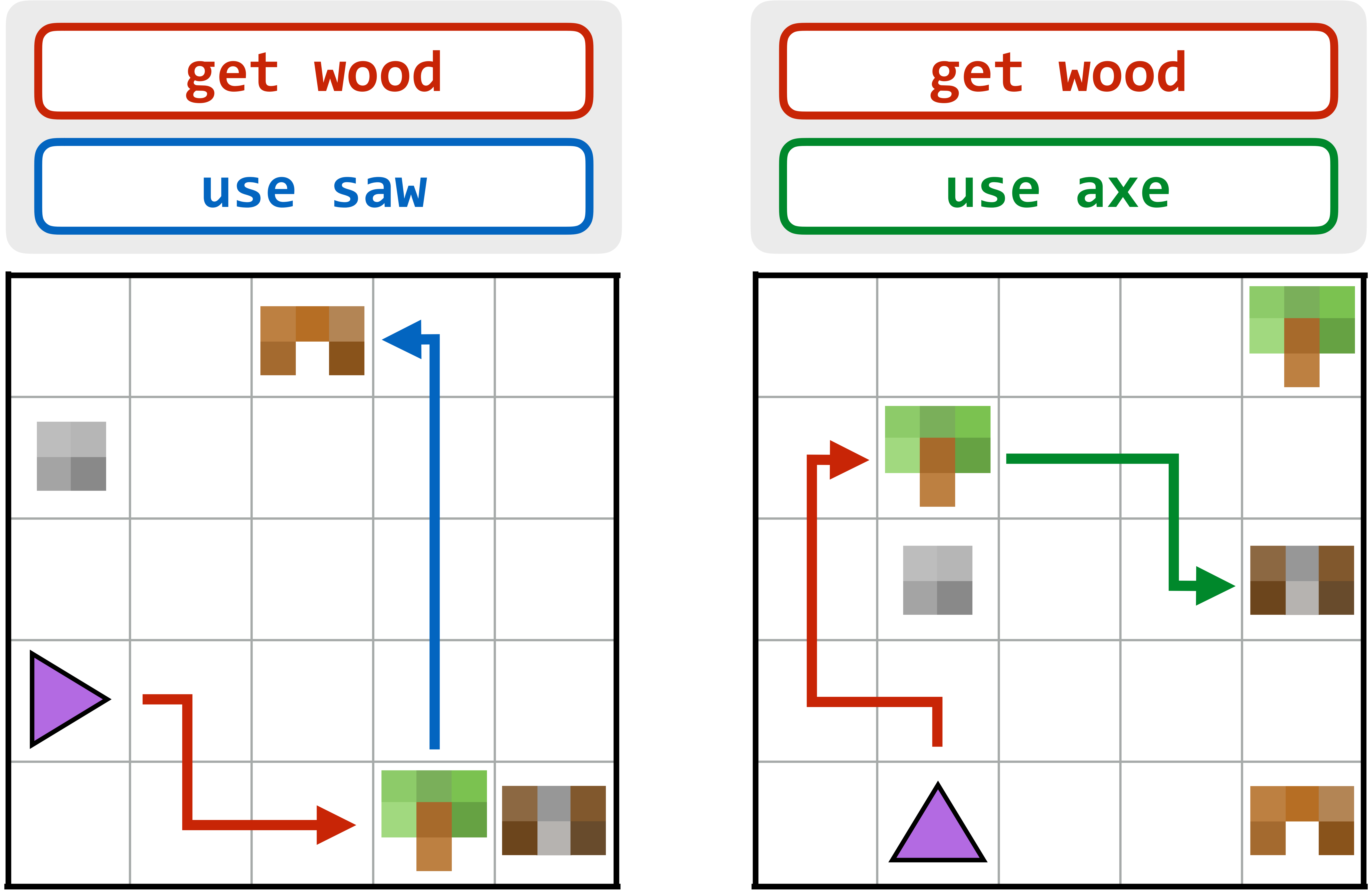
`make planks`



`make sticks`



Learning from sketches



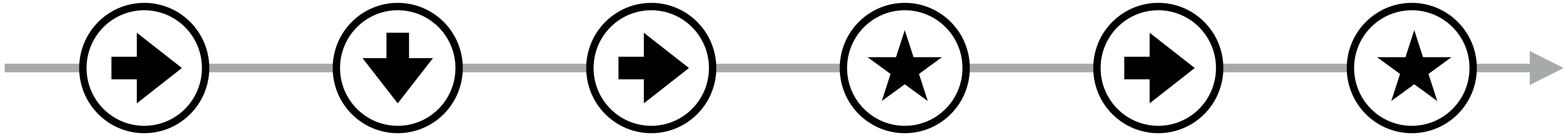
get wood

use saw

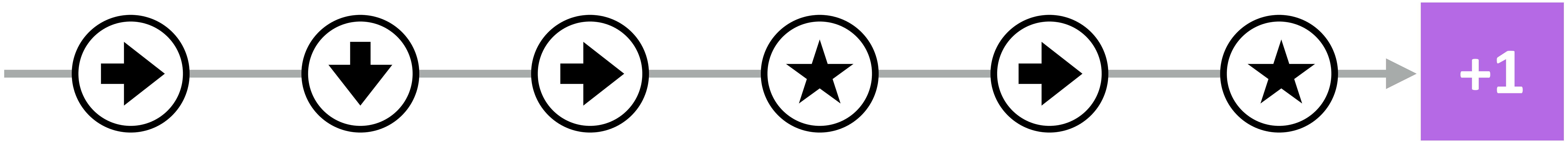
get wood

use axe

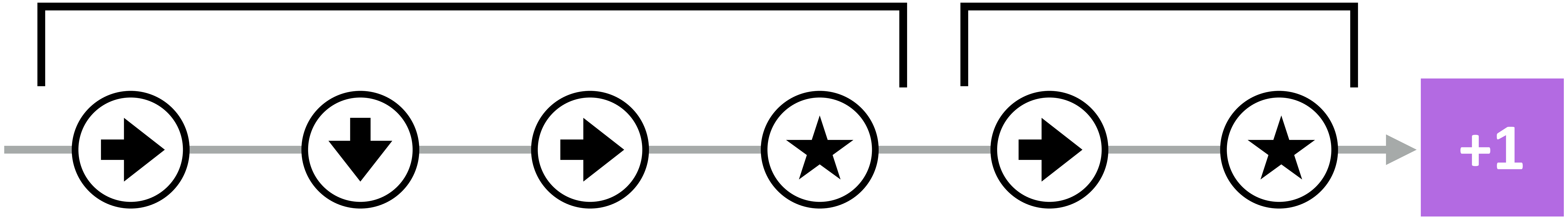
The options framework



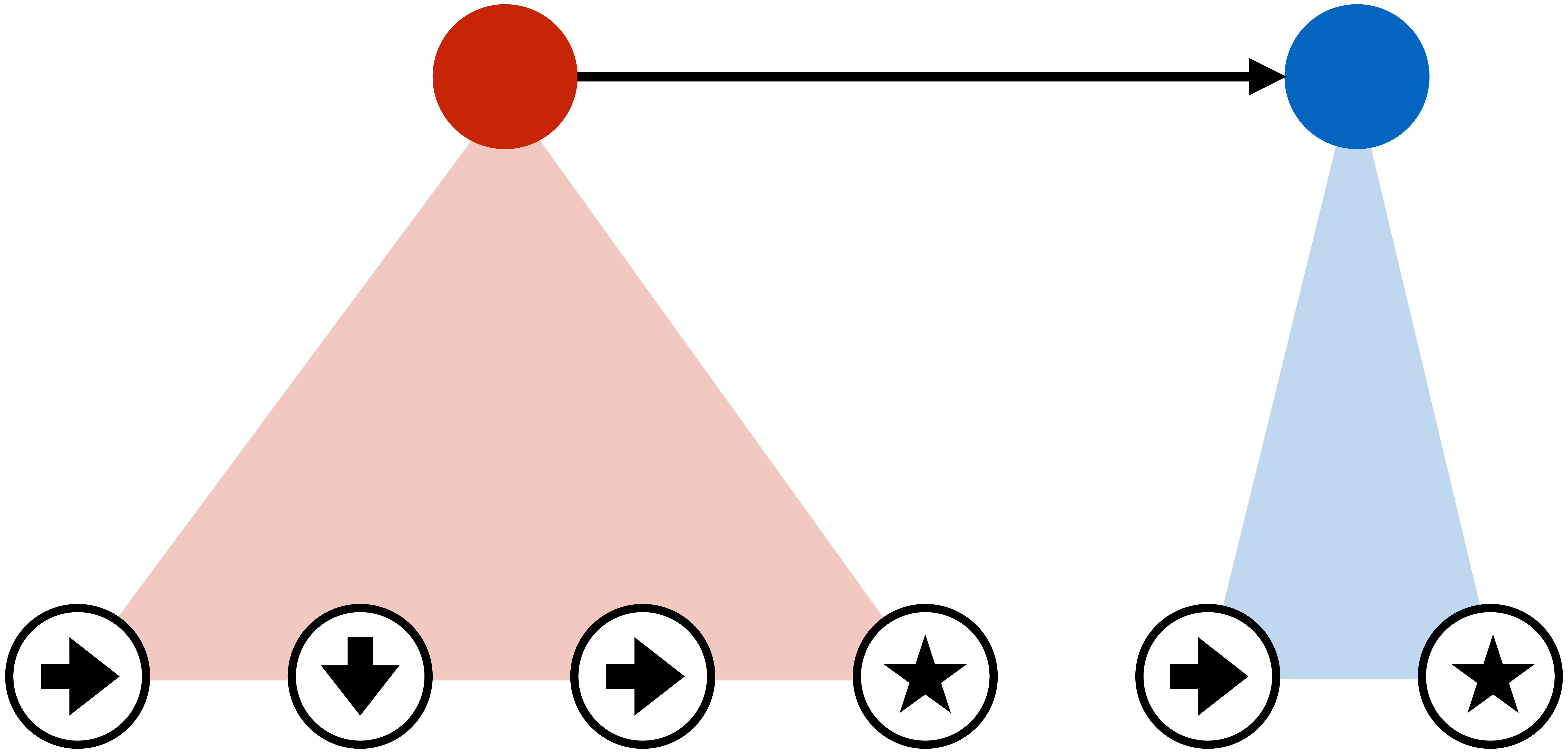
The options framework



The options framework

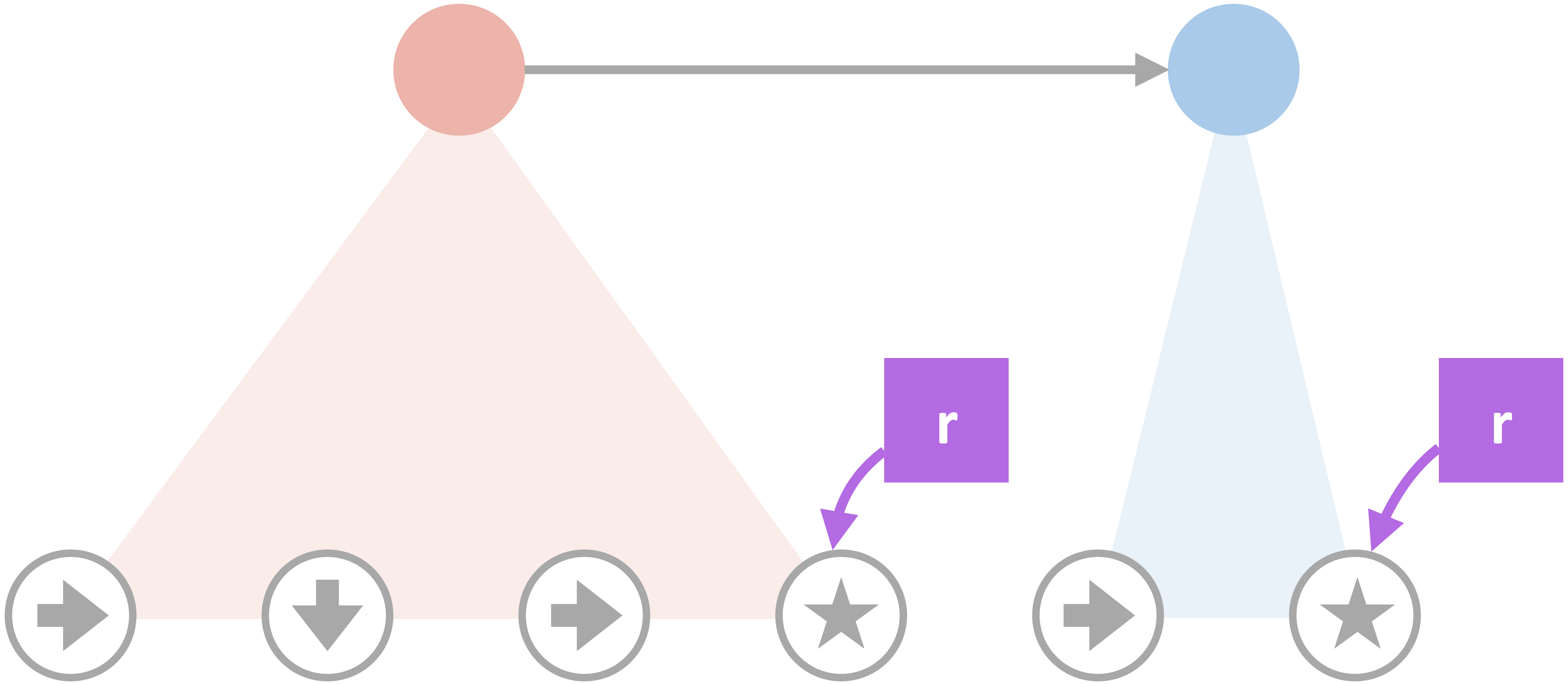


The options framework



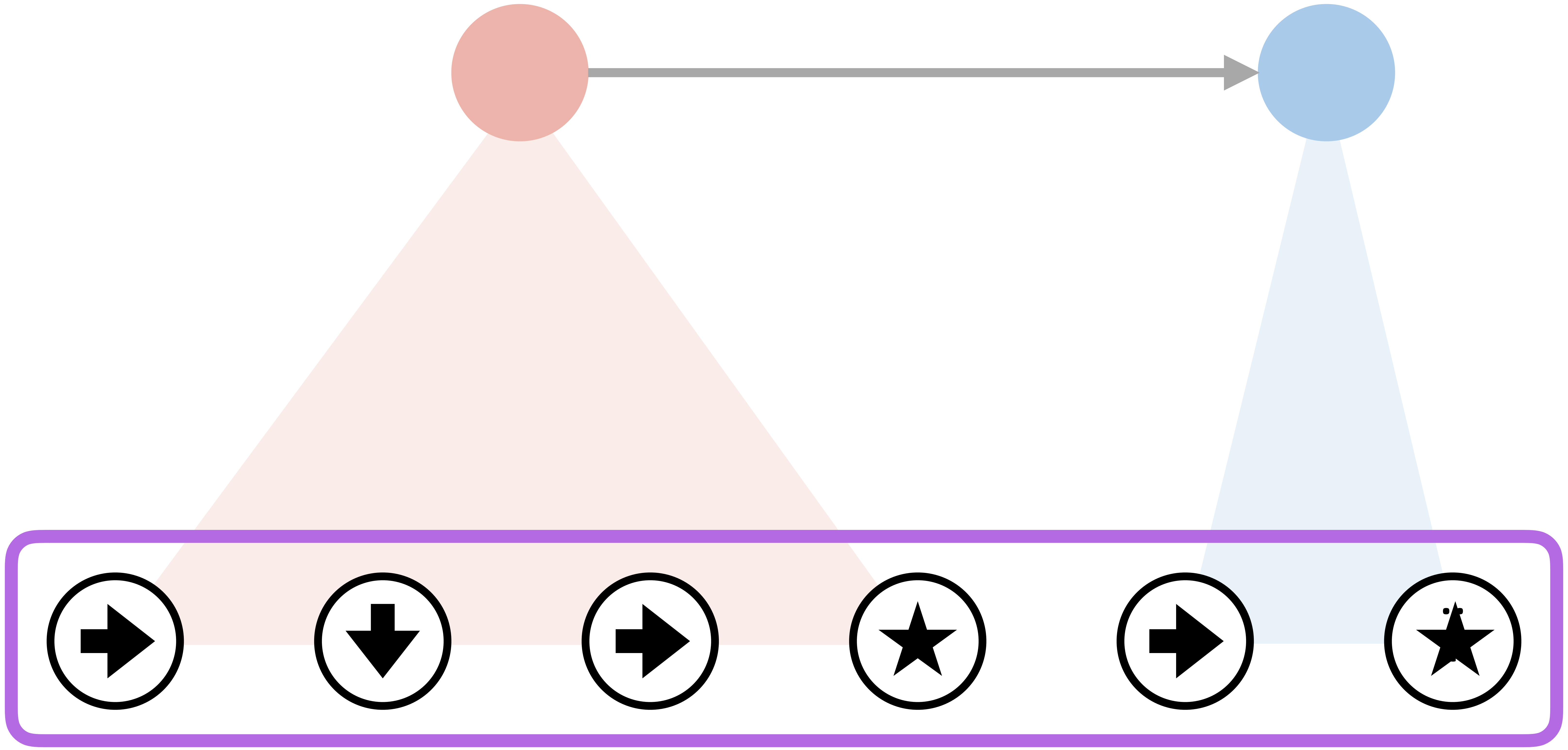
[Sutton et al. 99, Bacon & Precup 16]

Learning from intermediate rewards



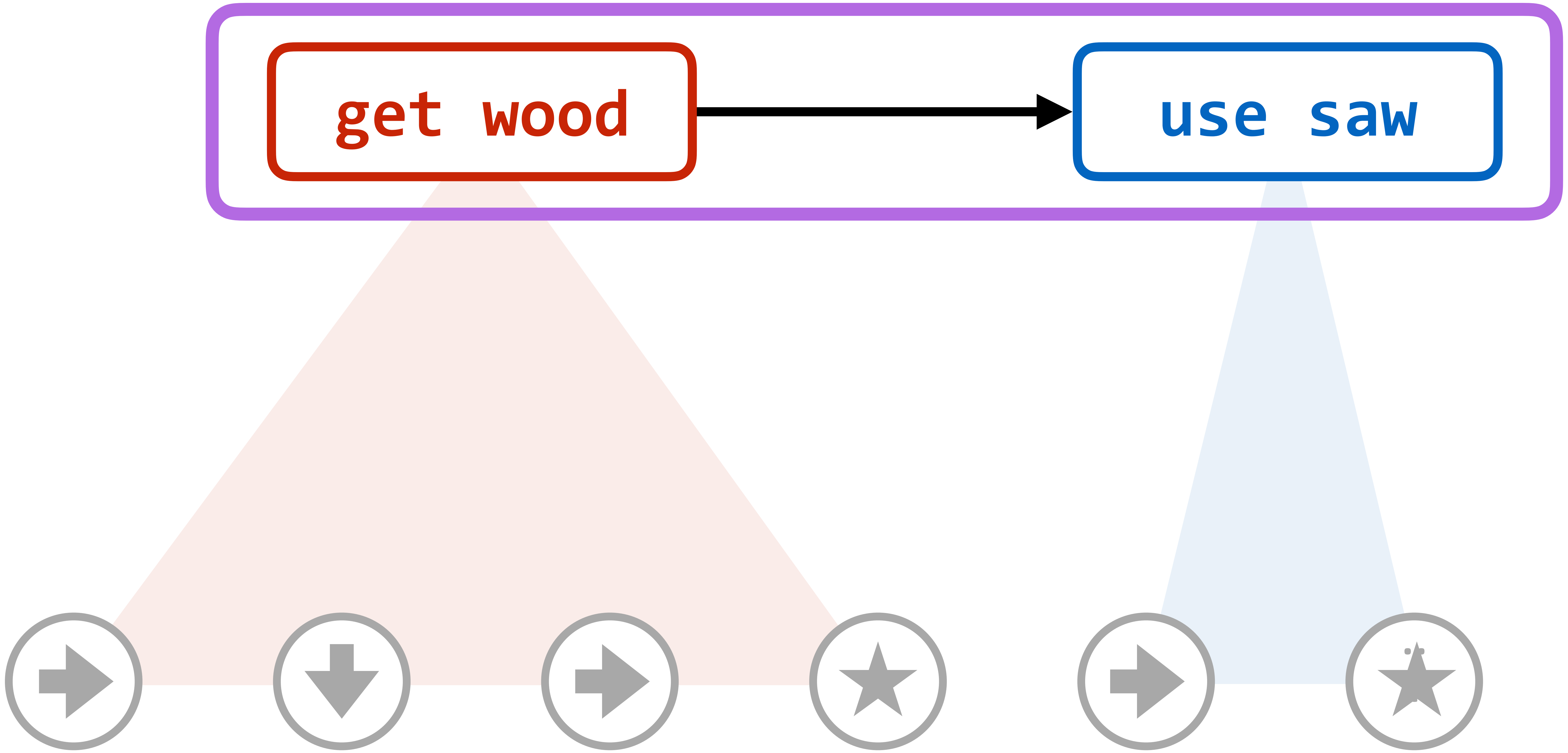
[Kearns & Singh 02, Kulkarni et al. 16]

Learning from demonstrations



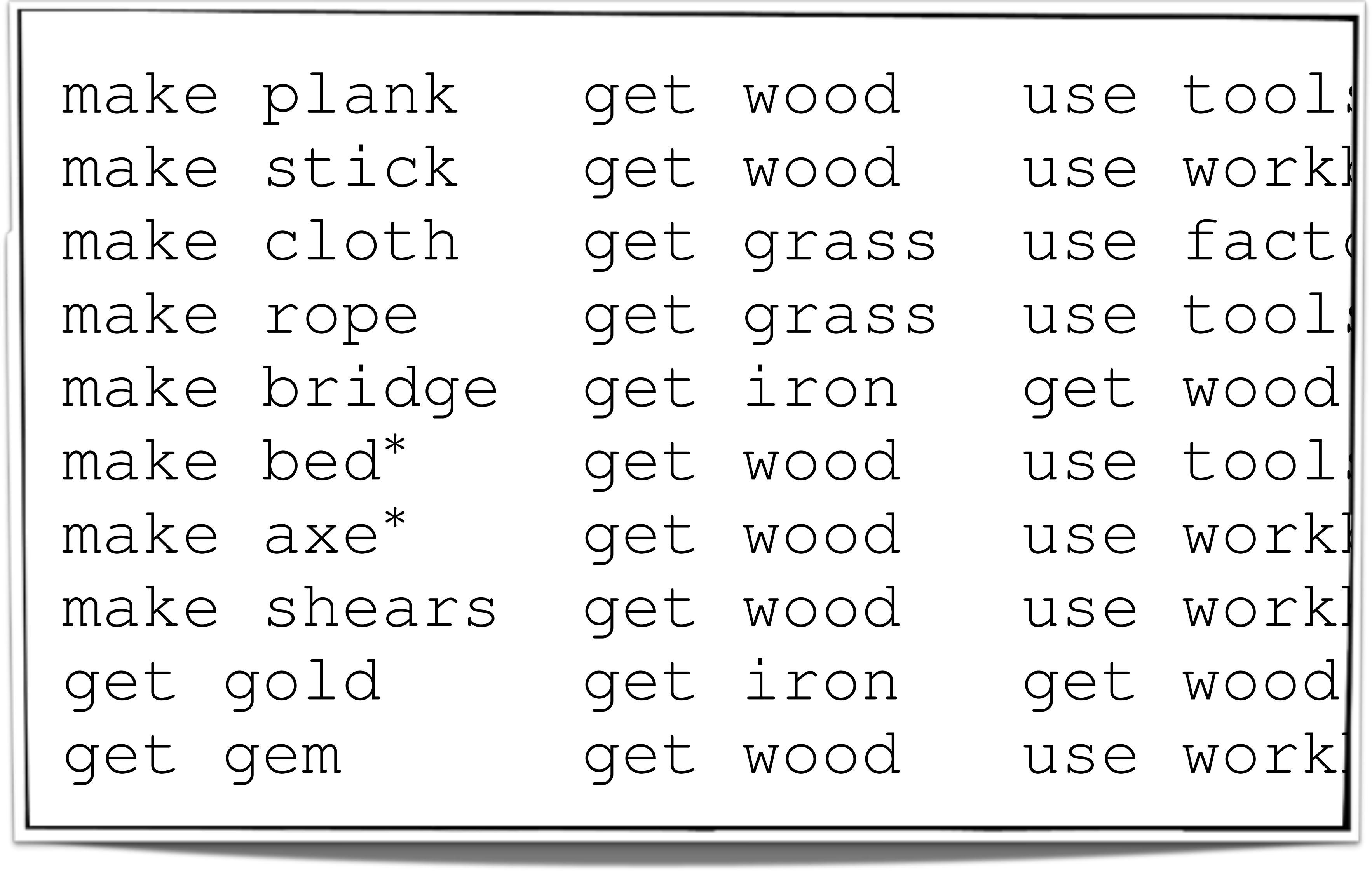
[Stolle & Precup 02, Fox & Krishnan et al. 16]

Learning from policy sketches



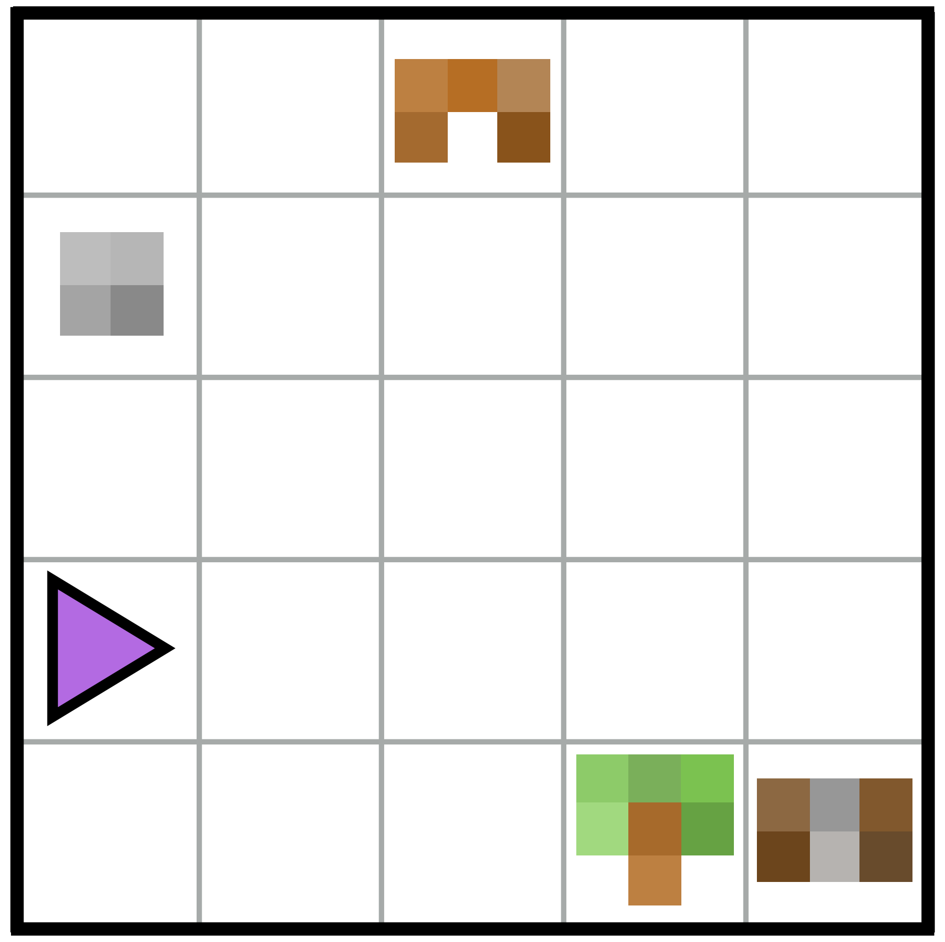
Why sketches?

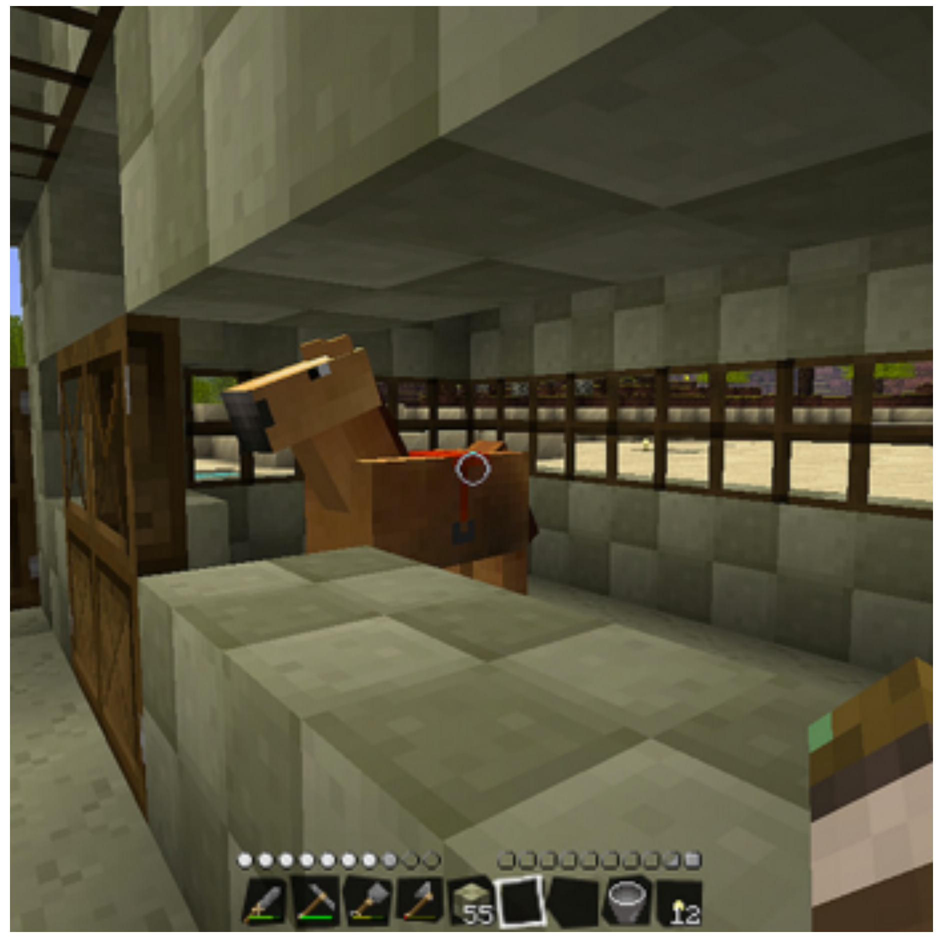
Easy to collect



```
make plank    get wood    use tool
make stick    get wood    use work
make cloth    get grass   use fact
make rope     get grass   use tool
make bridge   get iron    get wood
make bed*     get wood    use tool
make axe*     get wood    use work
make shears   get wood    use work
get gold      get iron    get wood
get gem       get wood    use work
```

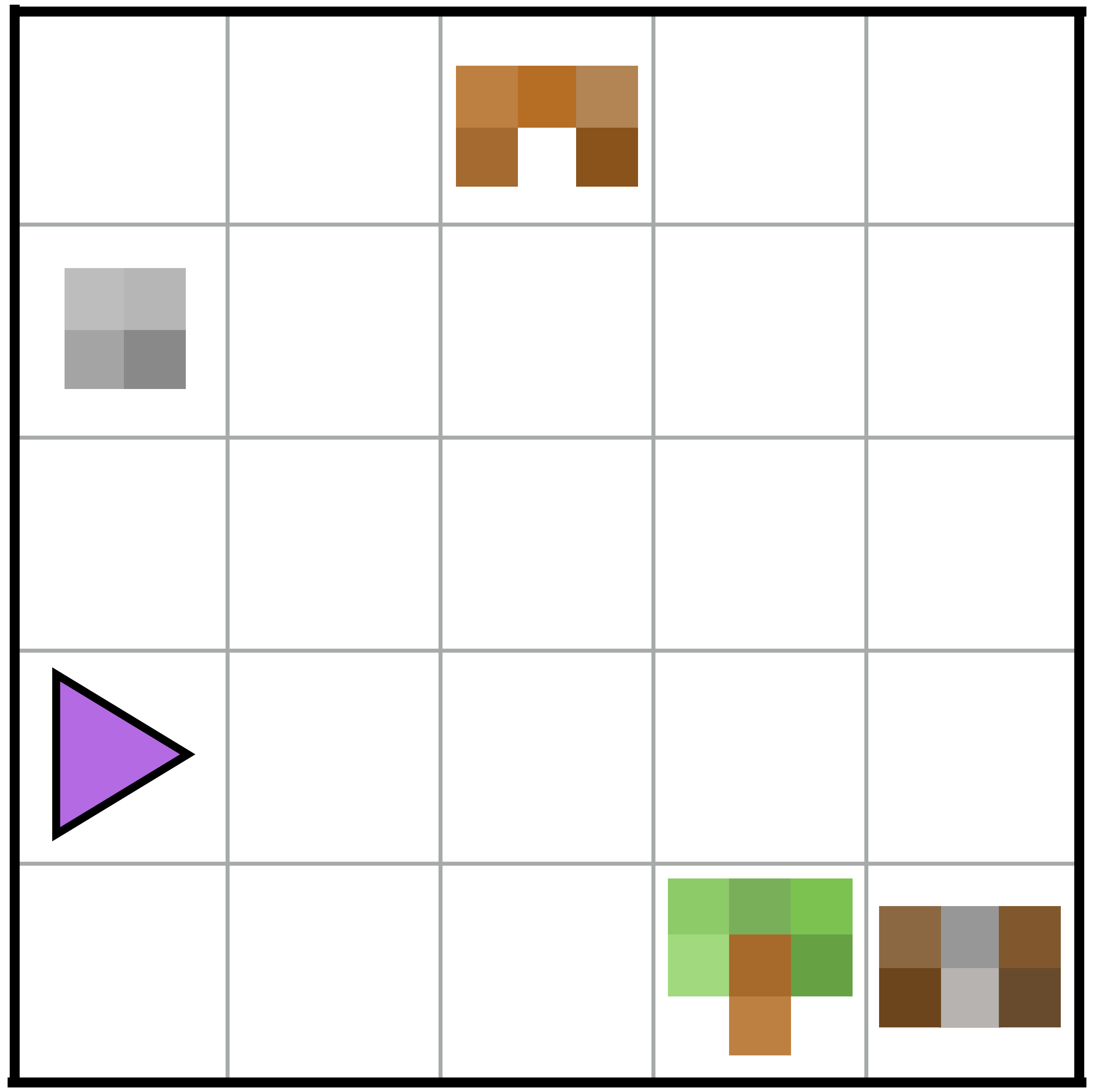
Portable





Learning from policy sketches

Learning from policy sketches

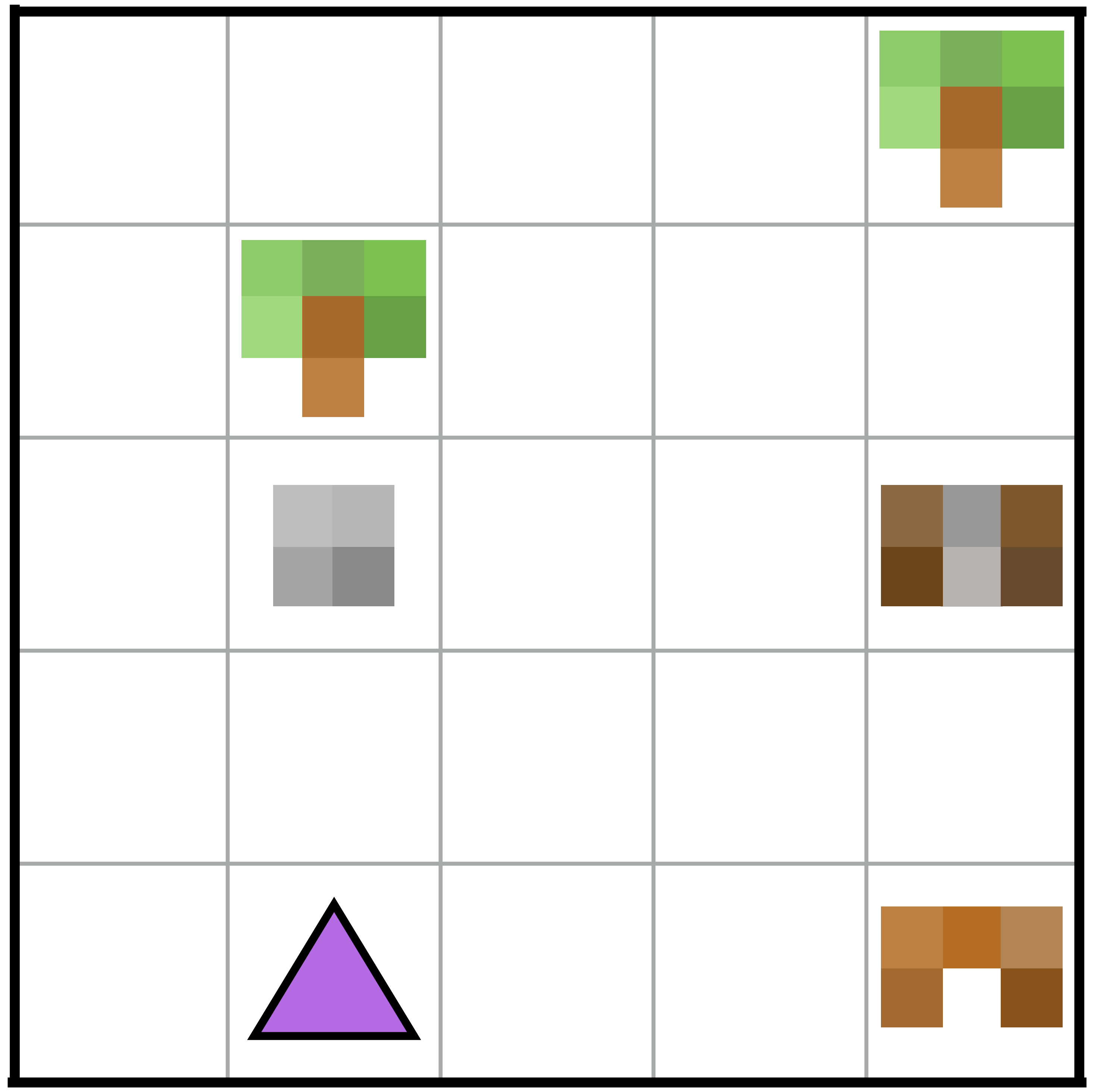


make planks

get wood

use saw

Learning from policy sketches

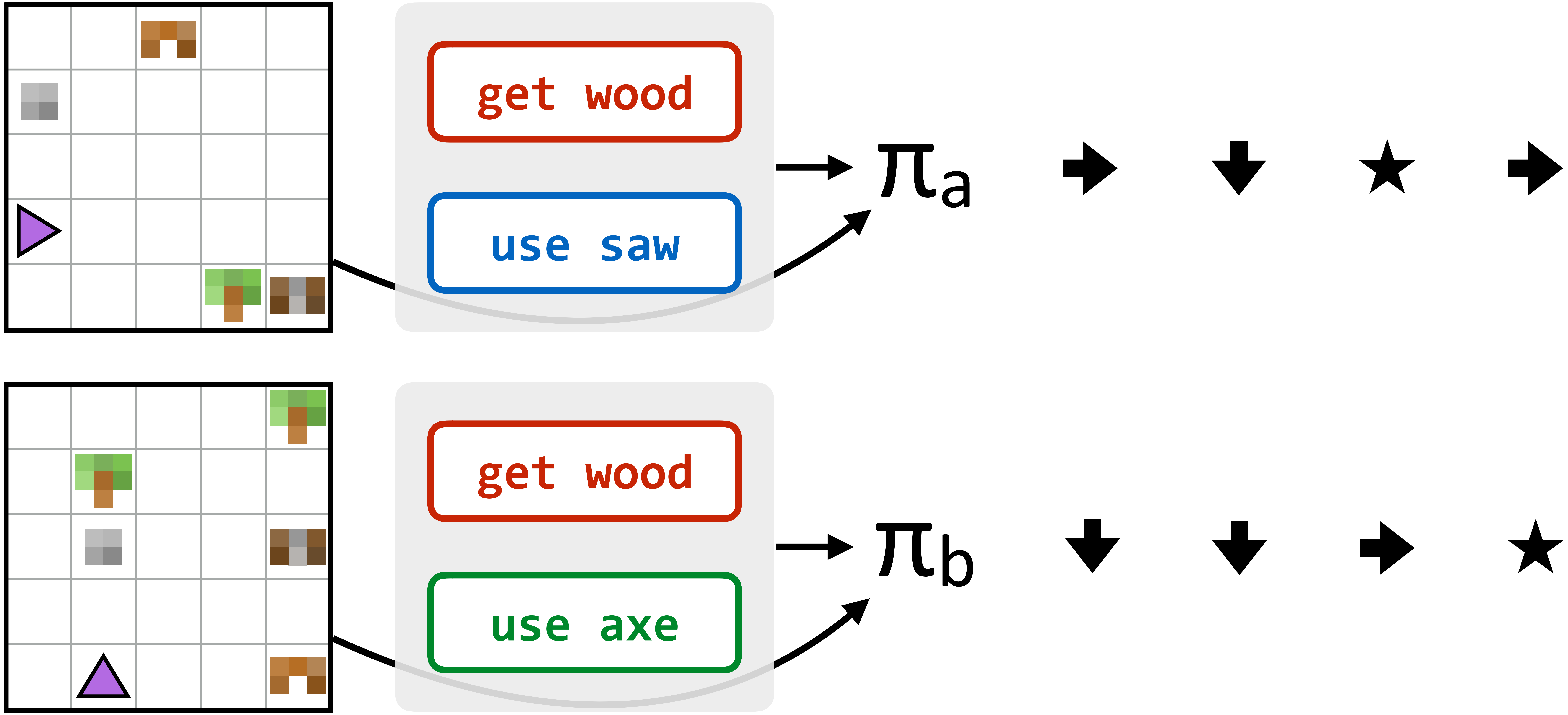


make sticks

get wood

use axe

Learning from policy sketches



get wood

use saw

π_a

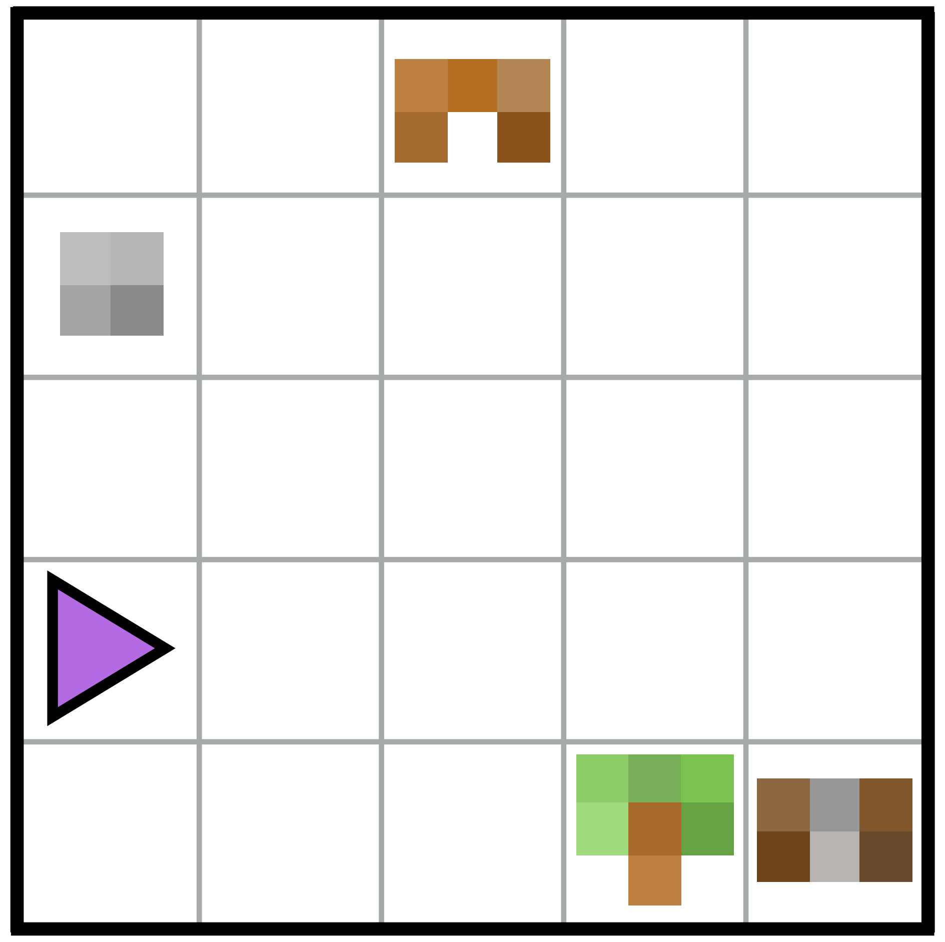
get wood

use axe

π_b

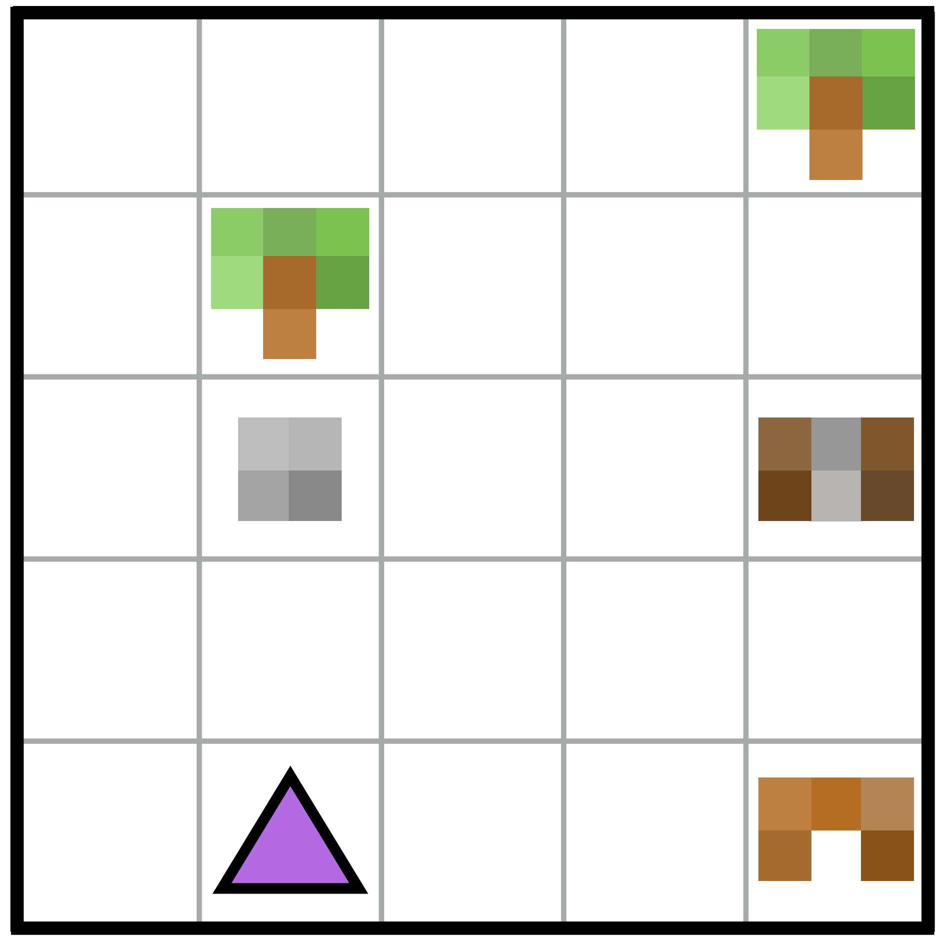
[e.g. Branavan et al. 09, Oh et al. 17, Hermann et al. 17]

Learning from policy sketches



get wood

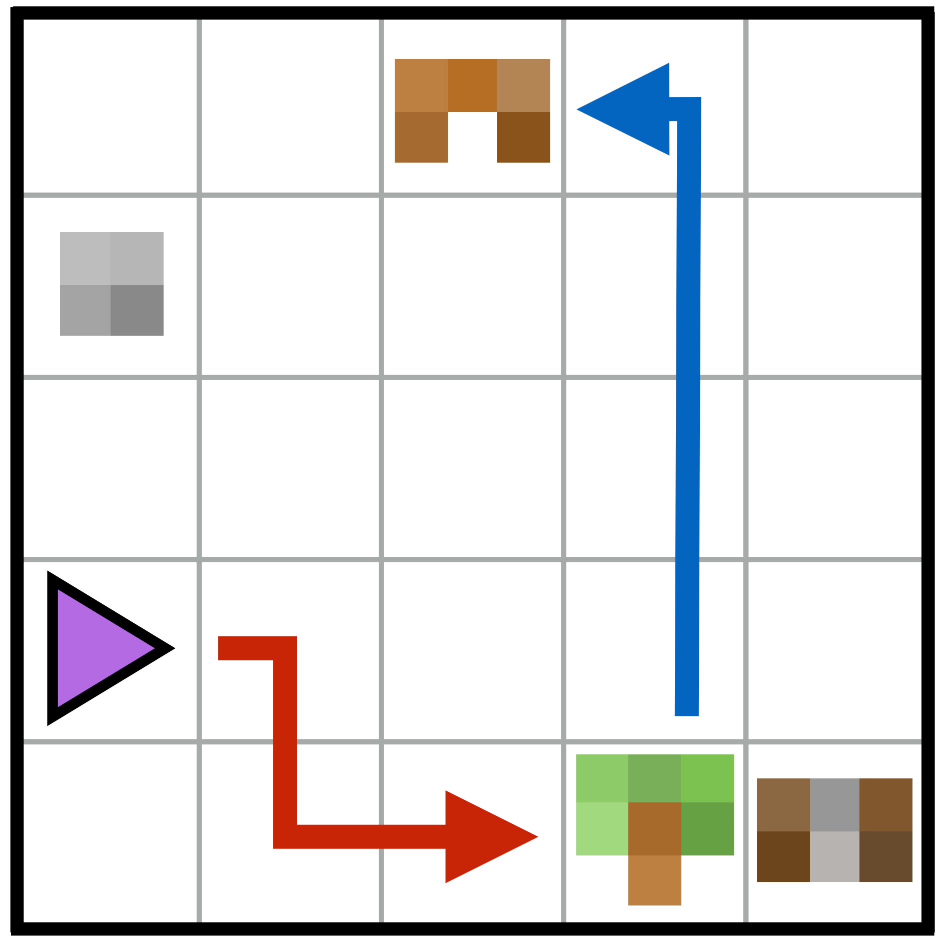
use saw



get wood

use axe

`

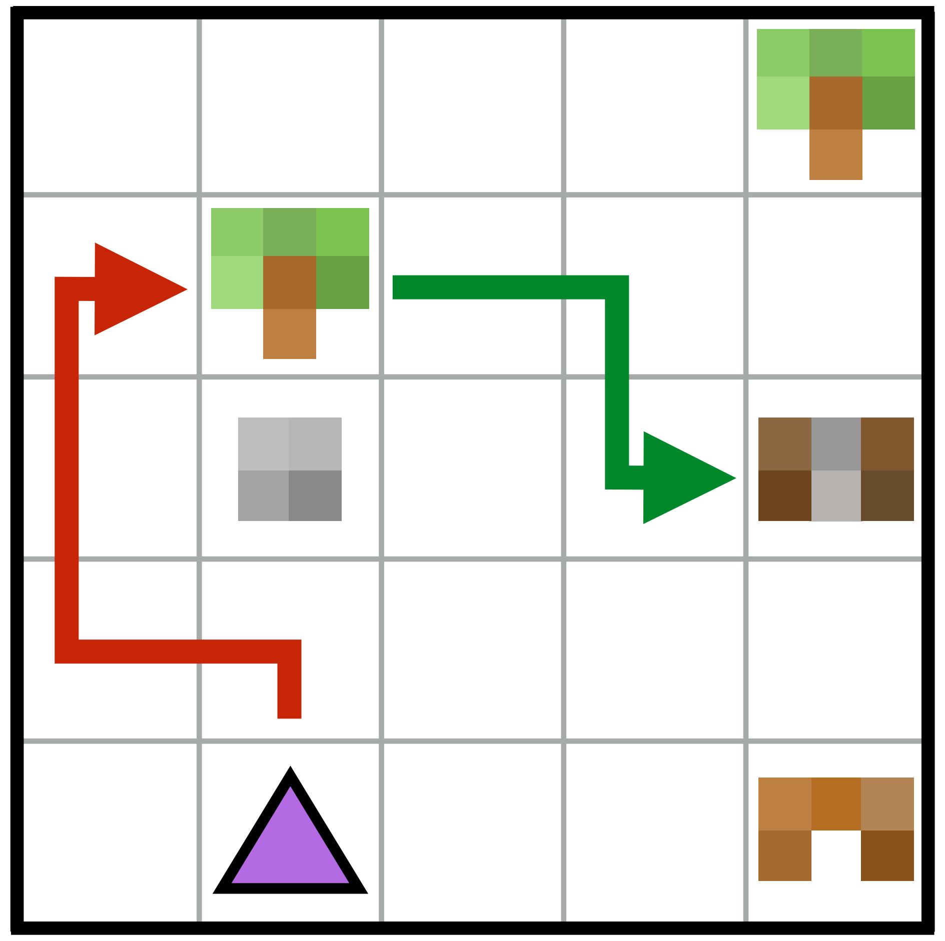


get wood

use saw

π_1

π_2



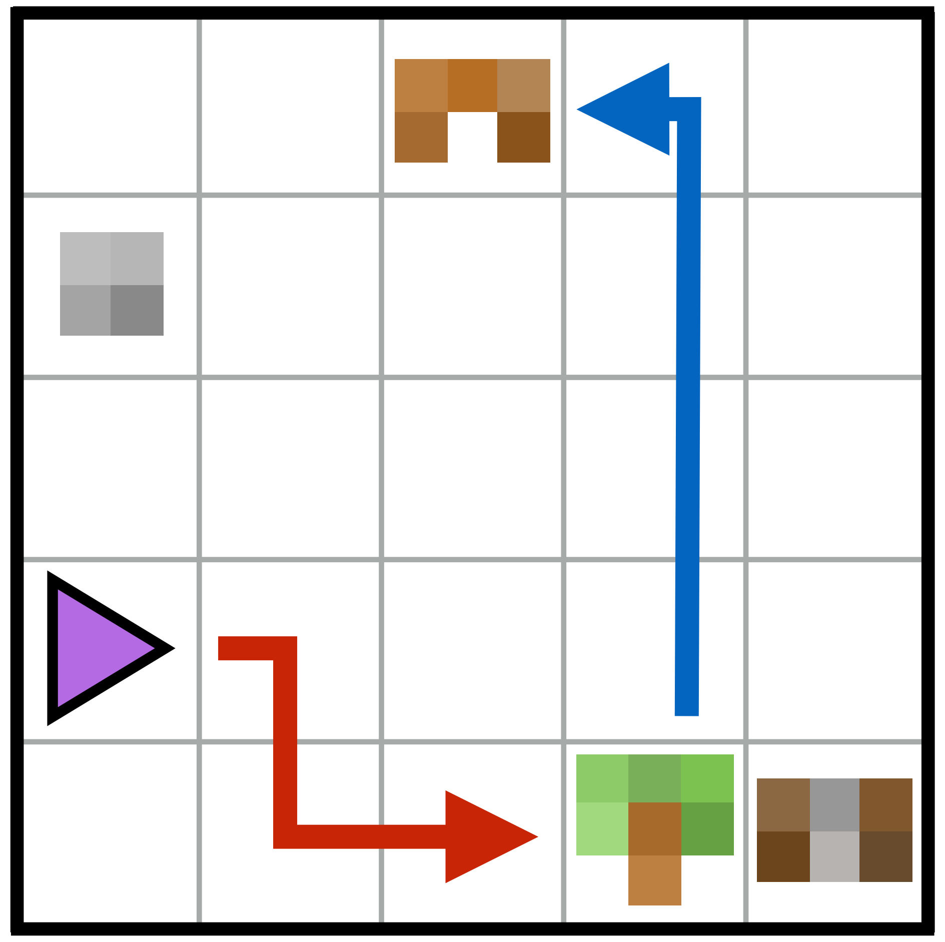
get wood

use axe

π_1

π_3

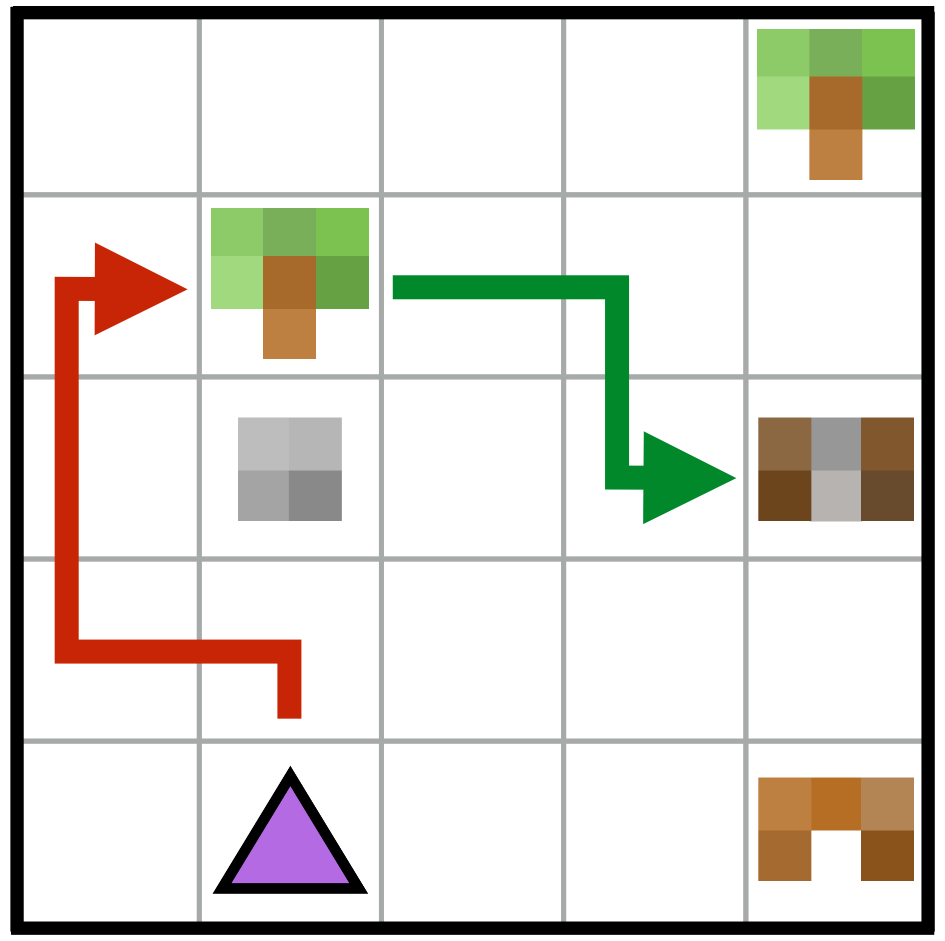
`



get wood   use saw

π_1 π_2

➡ ⬇ ⯃ ★ ➡ ⯃



get wood   use axe

π_1 π_3

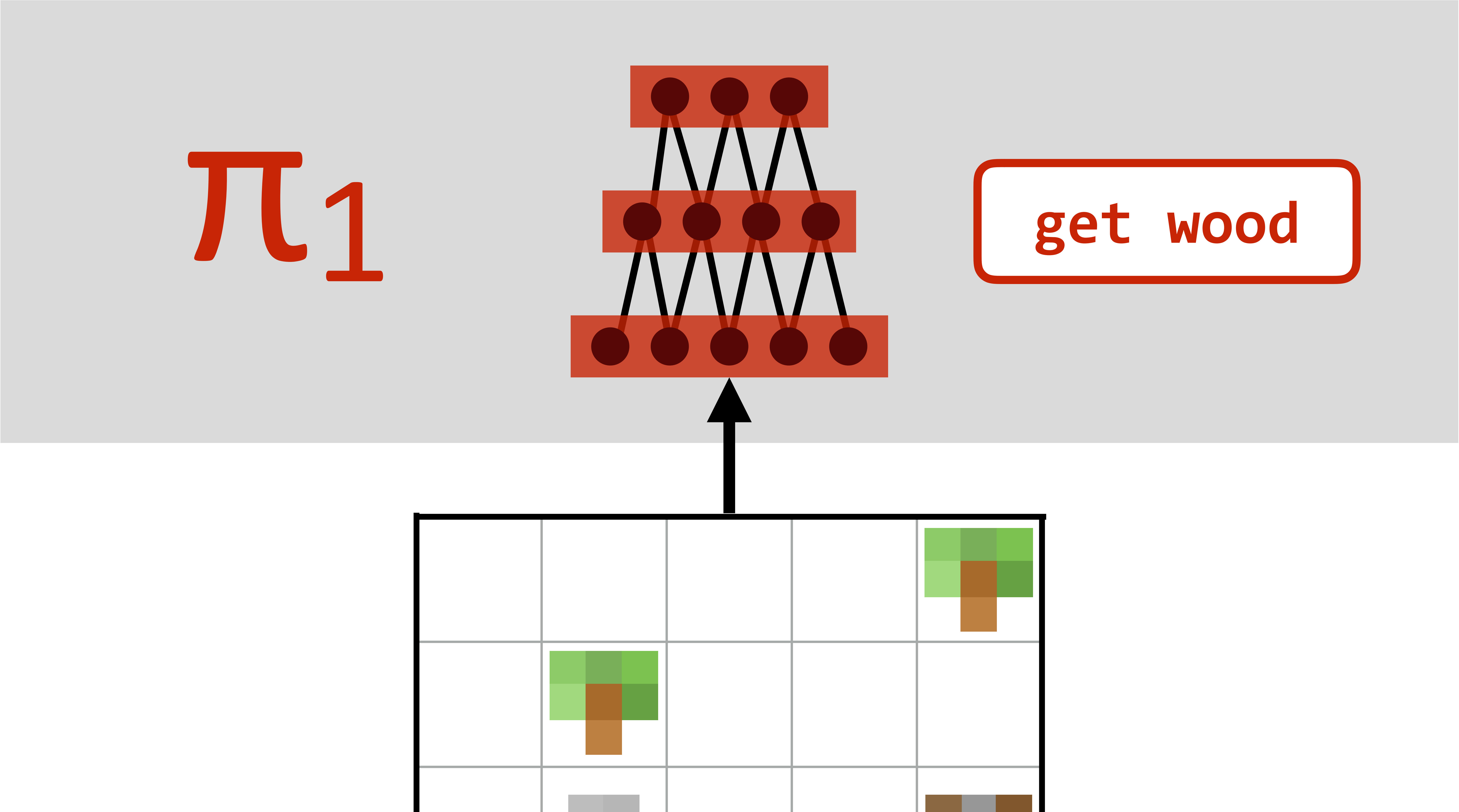
➡ ⬇ ★ ⯃ ➡ ➡

`

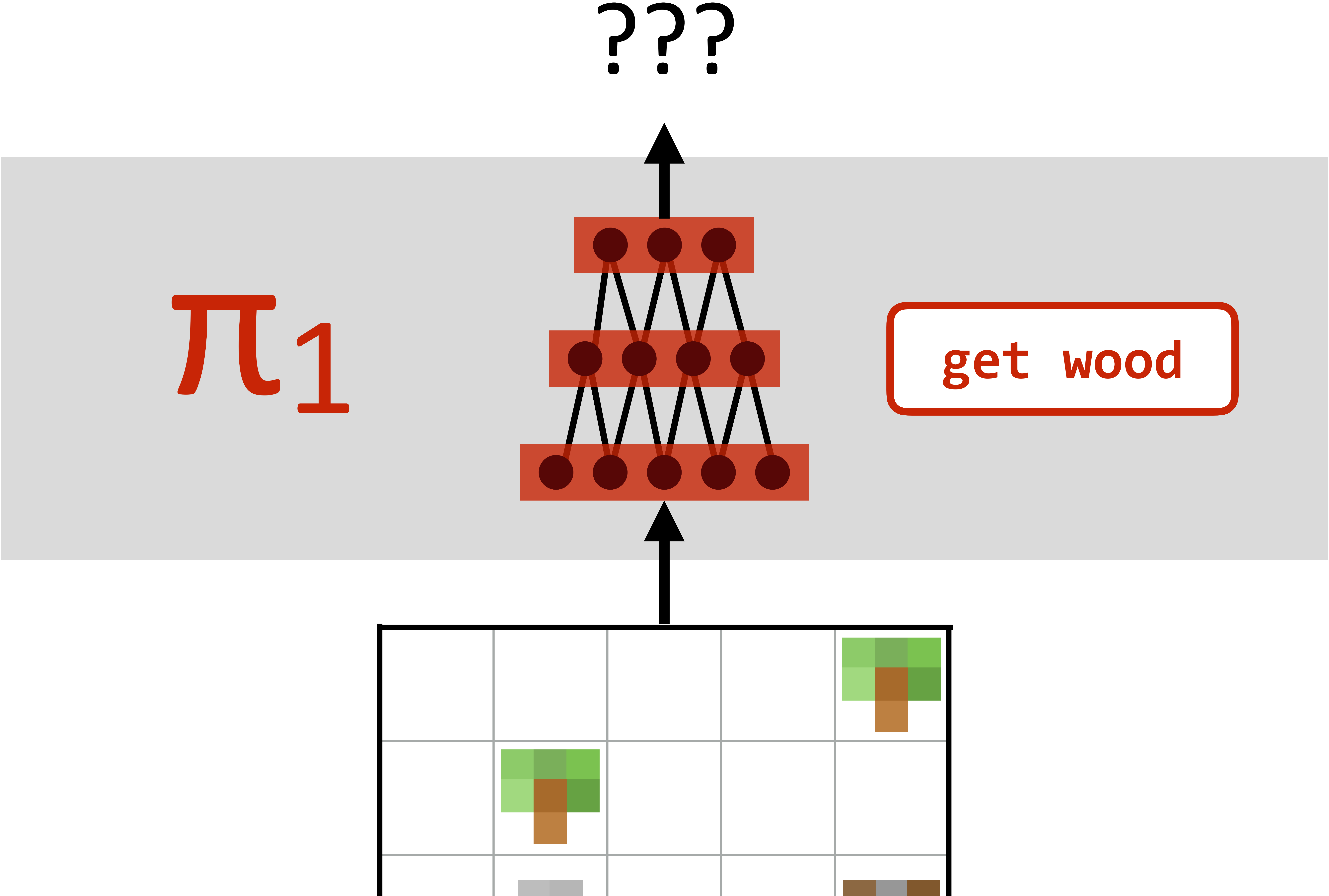
get wood

π_1

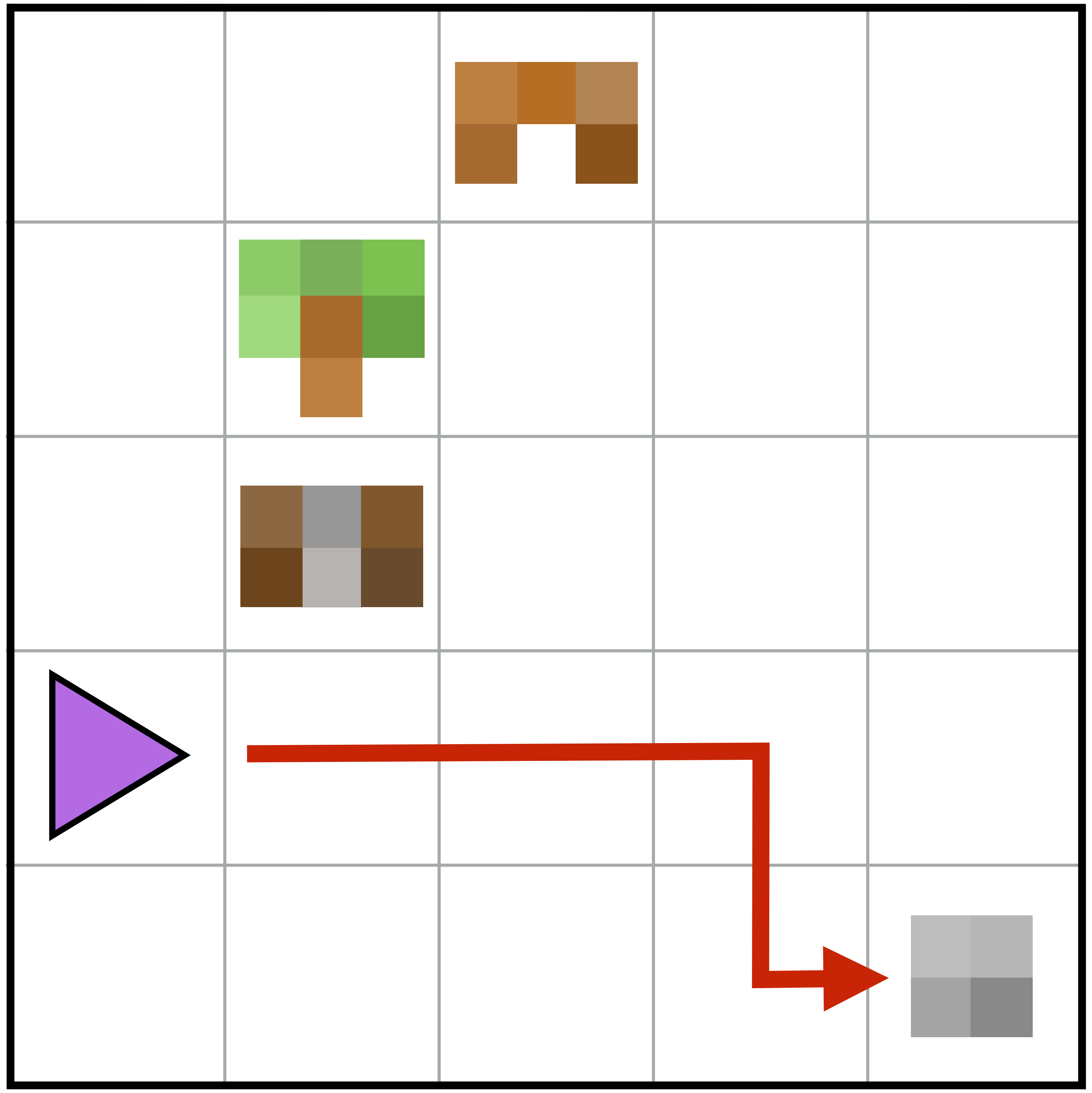
Policy representation



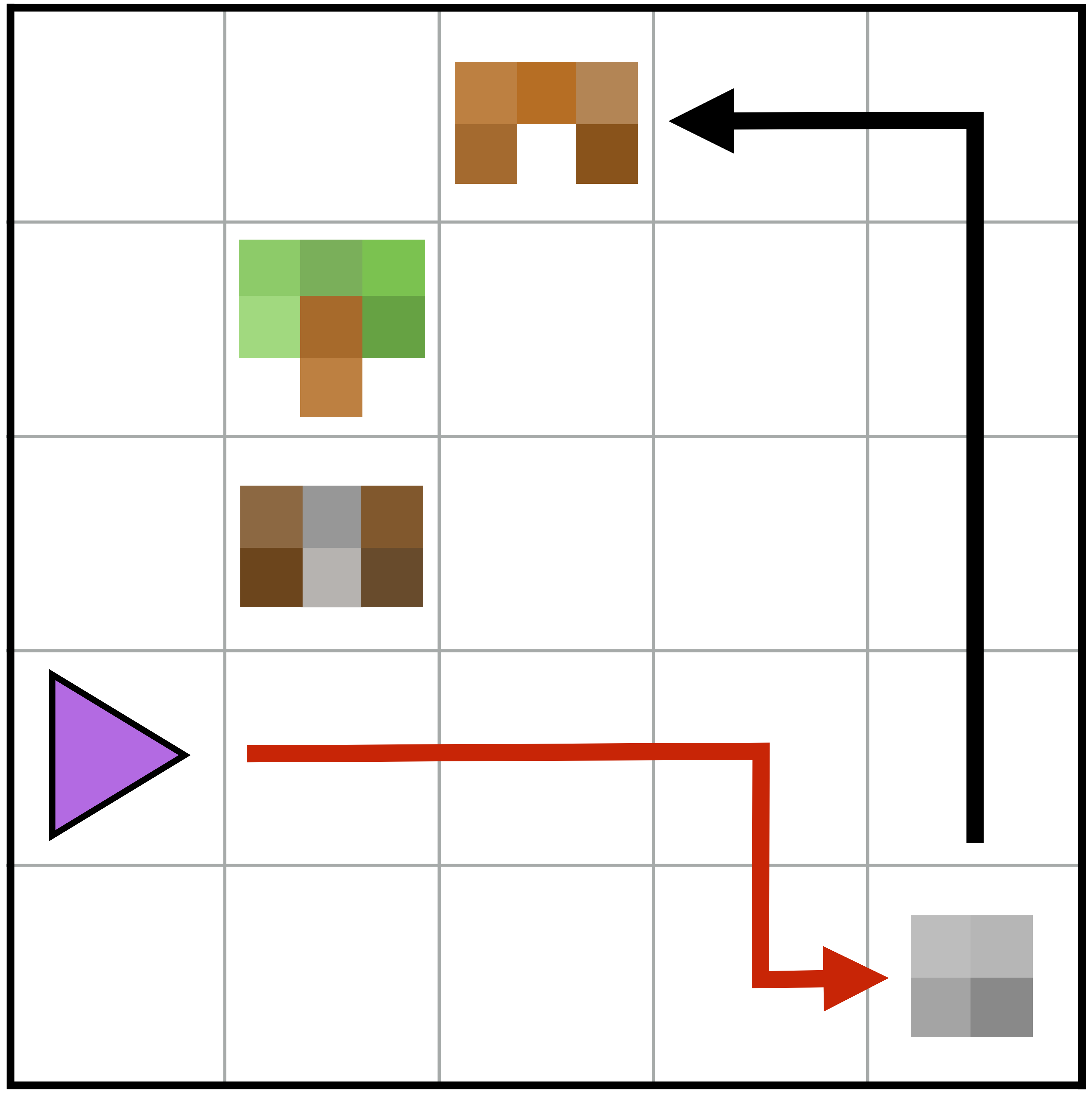
Policy representation



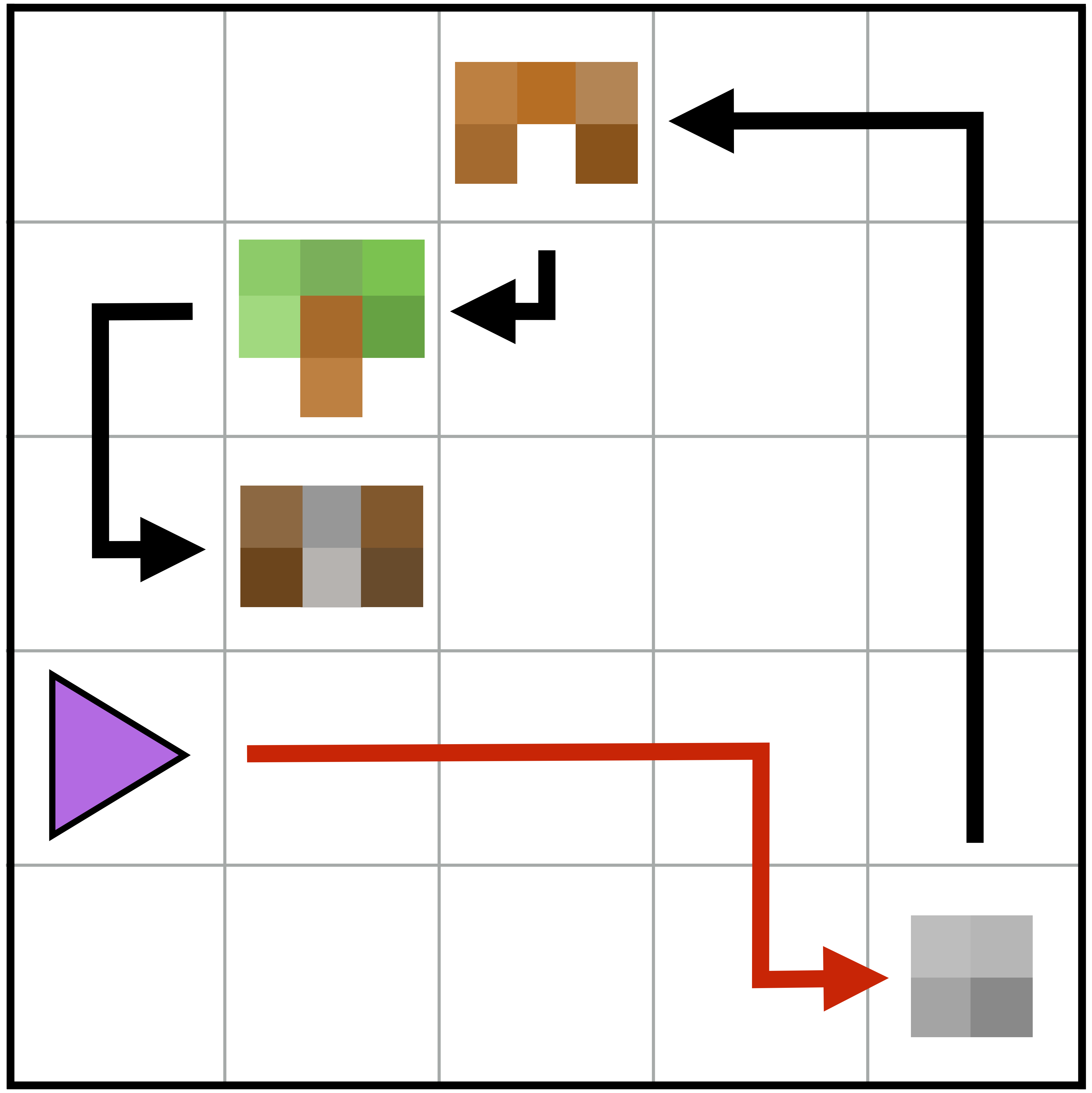
Policy representation



Policy representation

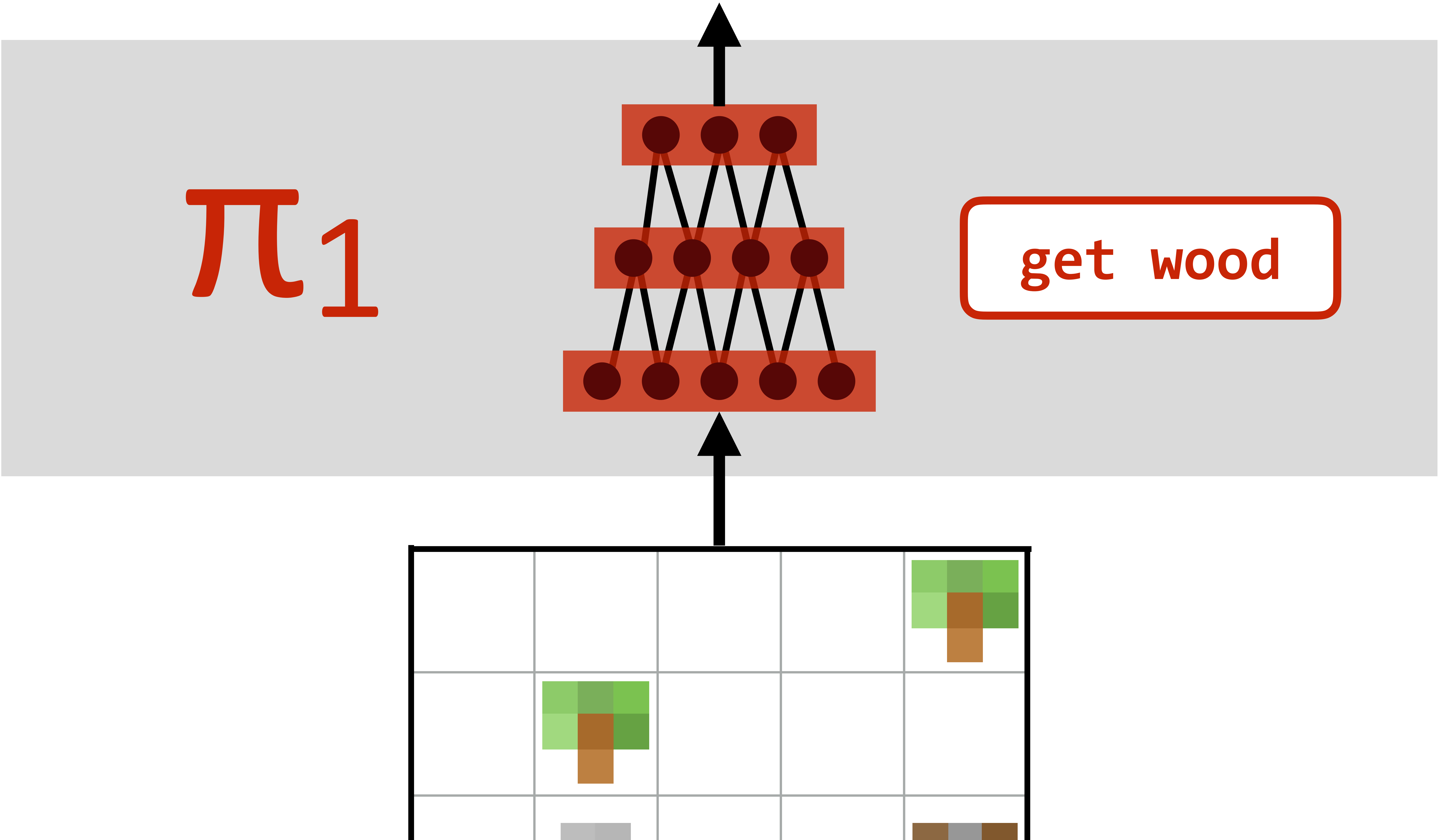


Policy representation



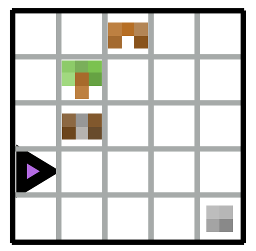
Policy representation

Action probabilities

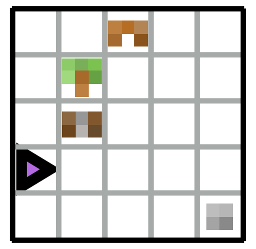


Policy search

action state reward baseline

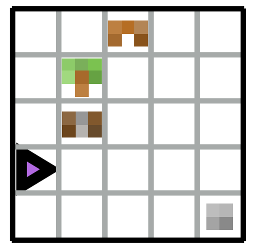
$\underset{\text{tasks}}{\sum}$ $\underset{\text{steps}}{\sum}$ $\Big(\nabla \log \pi(\rightarrow |$  $)\Big)$ $(r_t - b)$

Policy search

$\sum_{\text{tasks}} \sum_{\text{steps}} \left(\nabla \log \pi(\rightarrow |$  $)\right) (r_t - b)$

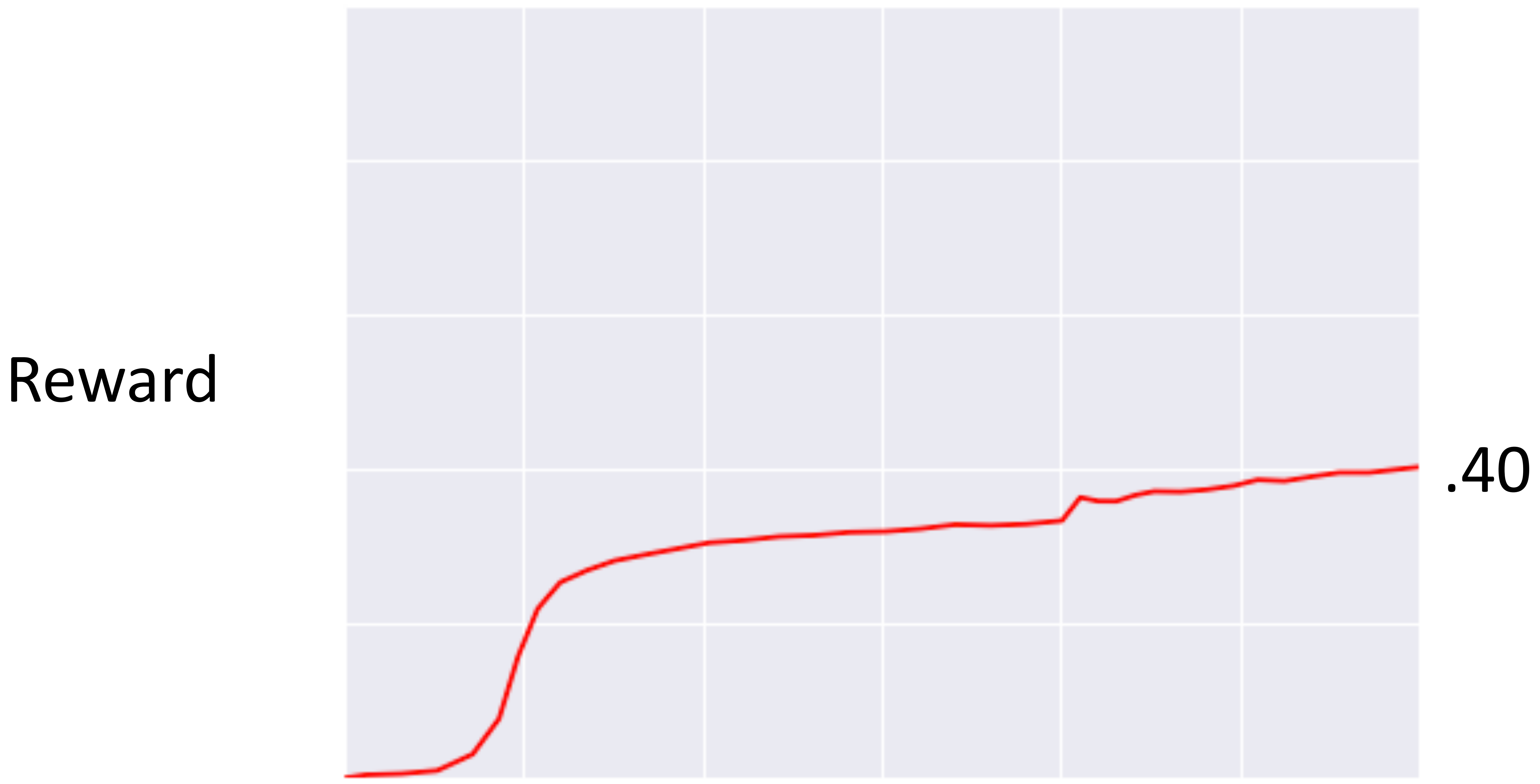
`get wood`

Policy search

$\sum_{\text{tasks}} \sum_{\text{steps}} \Big(\nabla \log \pi(\rightarrow \mid$  $)\Big)\ (r_t - b)$

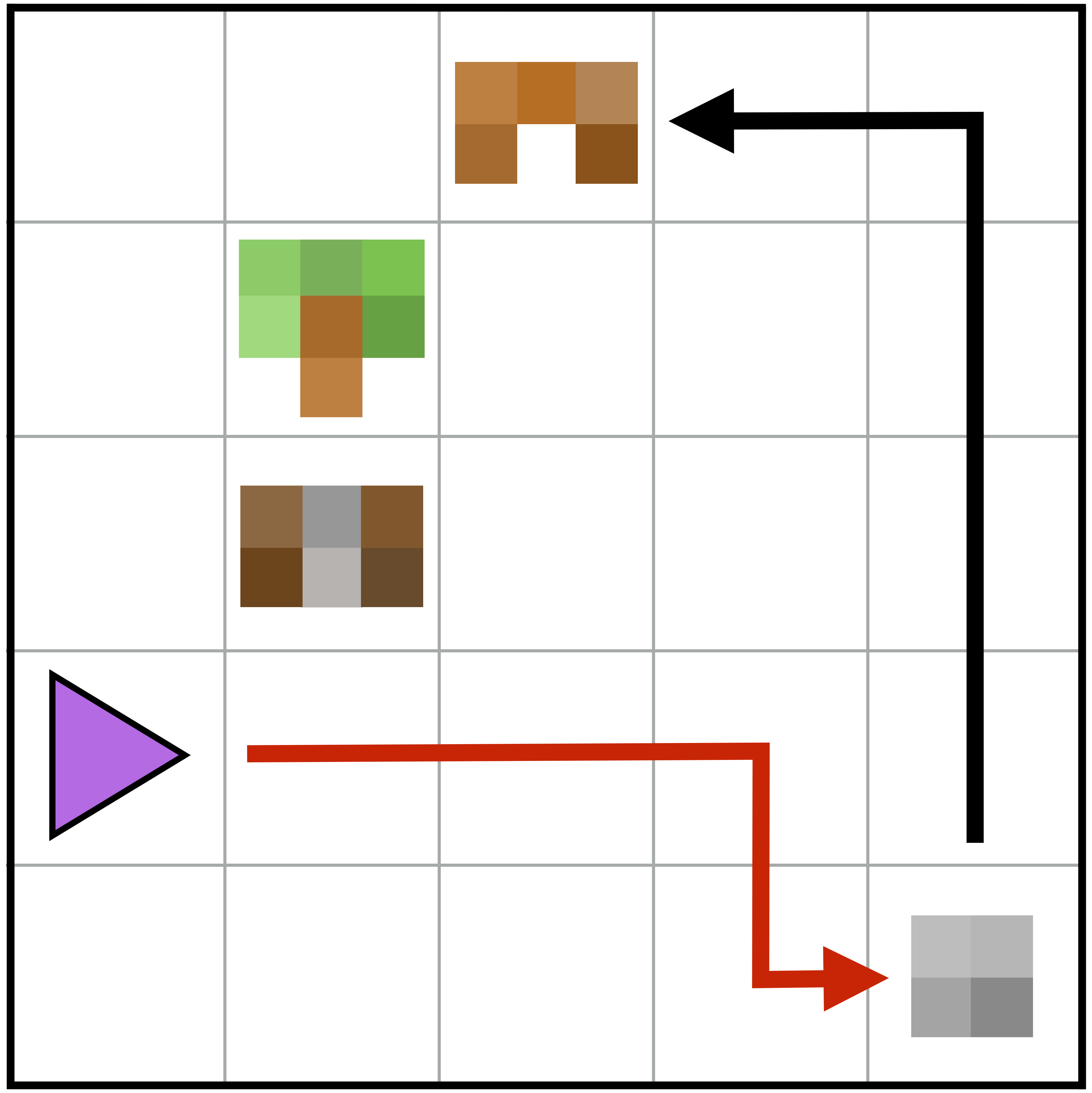
use axe

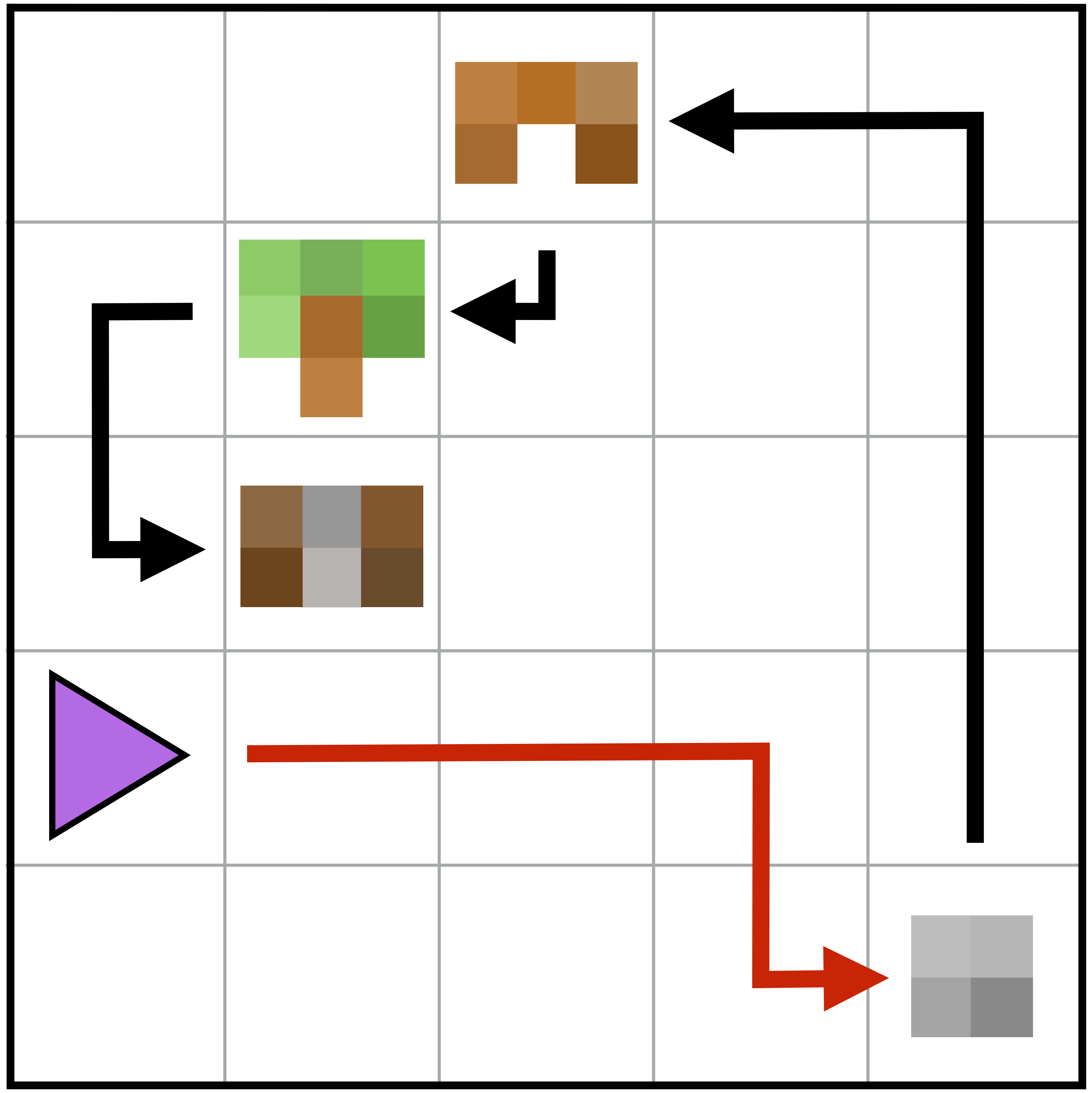
Policy search



$$\sum_{\text{tasks}} \; \sum_{\text{steps}} \left(\nabla \log \boxed{\texttt{SUBPOLICY}}\right) (r_t - b)$$

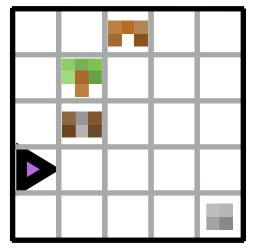
Improving policy search





Improving policy search

action state reward baseline

$\sum_{\text{tasks}} \sum_{\text{steps}} (\nabla \log \pi(\rightarrow |$  $)) \boxed{(r_t - b)}$

Improving policy search

$$(\nabla \log \boxed{\texttt{use saw}})(r_t - \boxed{\texttt{make planks}})$$

$$(\nabla \log \boxed{\texttt{use axe}})(r_t - \boxed{\texttt{make planks}})$$

$$(\nabla \log \boxed{\texttt{get wood}})(r_t - \boxed{\texttt{make planks}})$$

$$(\nabla \log \boxed{\texttt{get iron}})(r_t - \boxed{\texttt{make planks}})$$

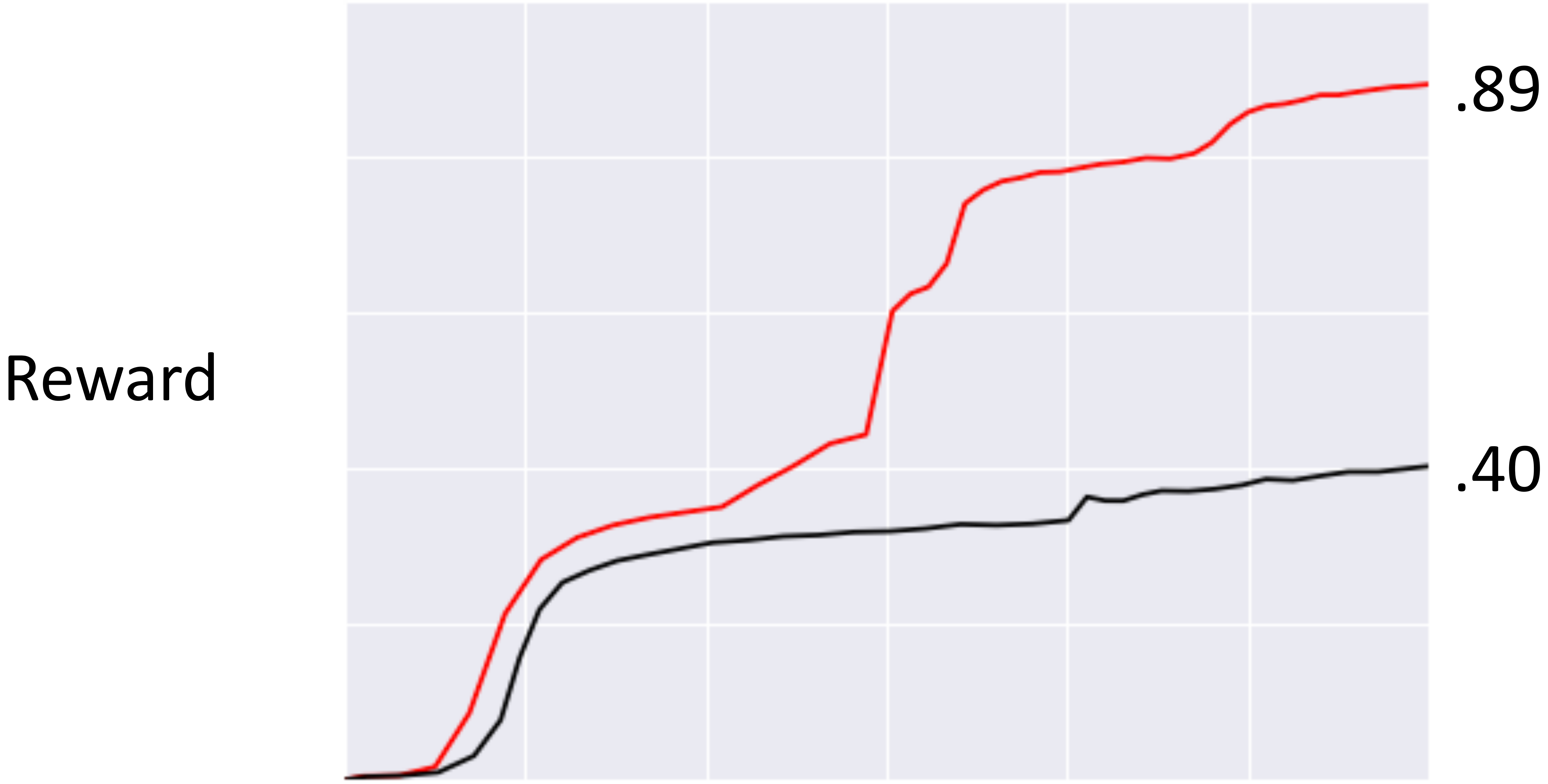
$$(\nabla \log \boxed{\texttt{use saw}})(r_t - \boxed{\texttt{make nails}})$$

$$(\nabla \log \boxed{\texttt{use axe}})(r_t - \boxed{\texttt{make nails}})$$

$$(\nabla \log \boxed{\texttt{get wood}})(r_t - \boxed{\texttt{make nails}})$$

$$(\nabla \log \boxed{\texttt{get iron}})(r_t - \boxed{\texttt{make nails}})$$

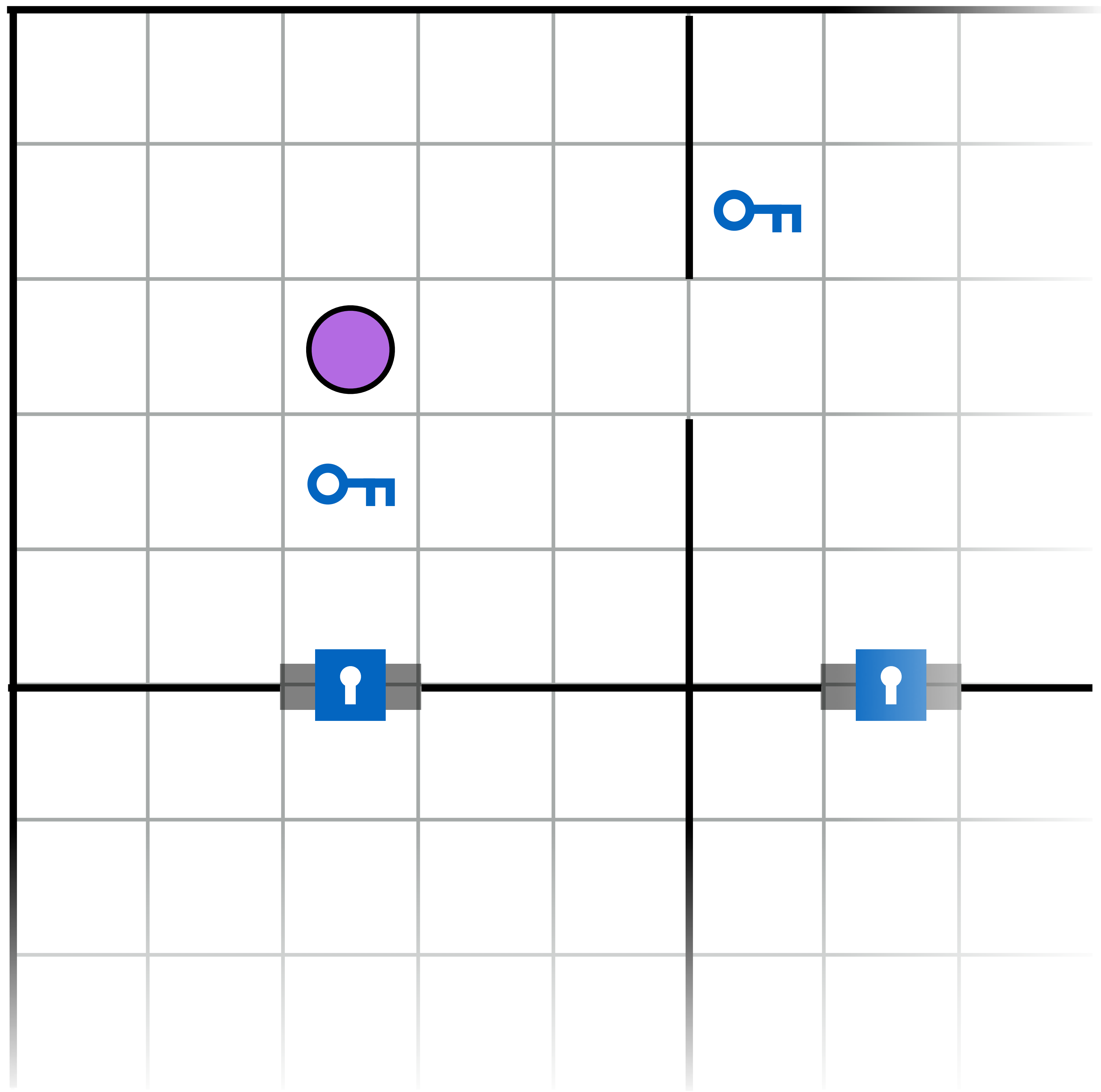
Improving policy search



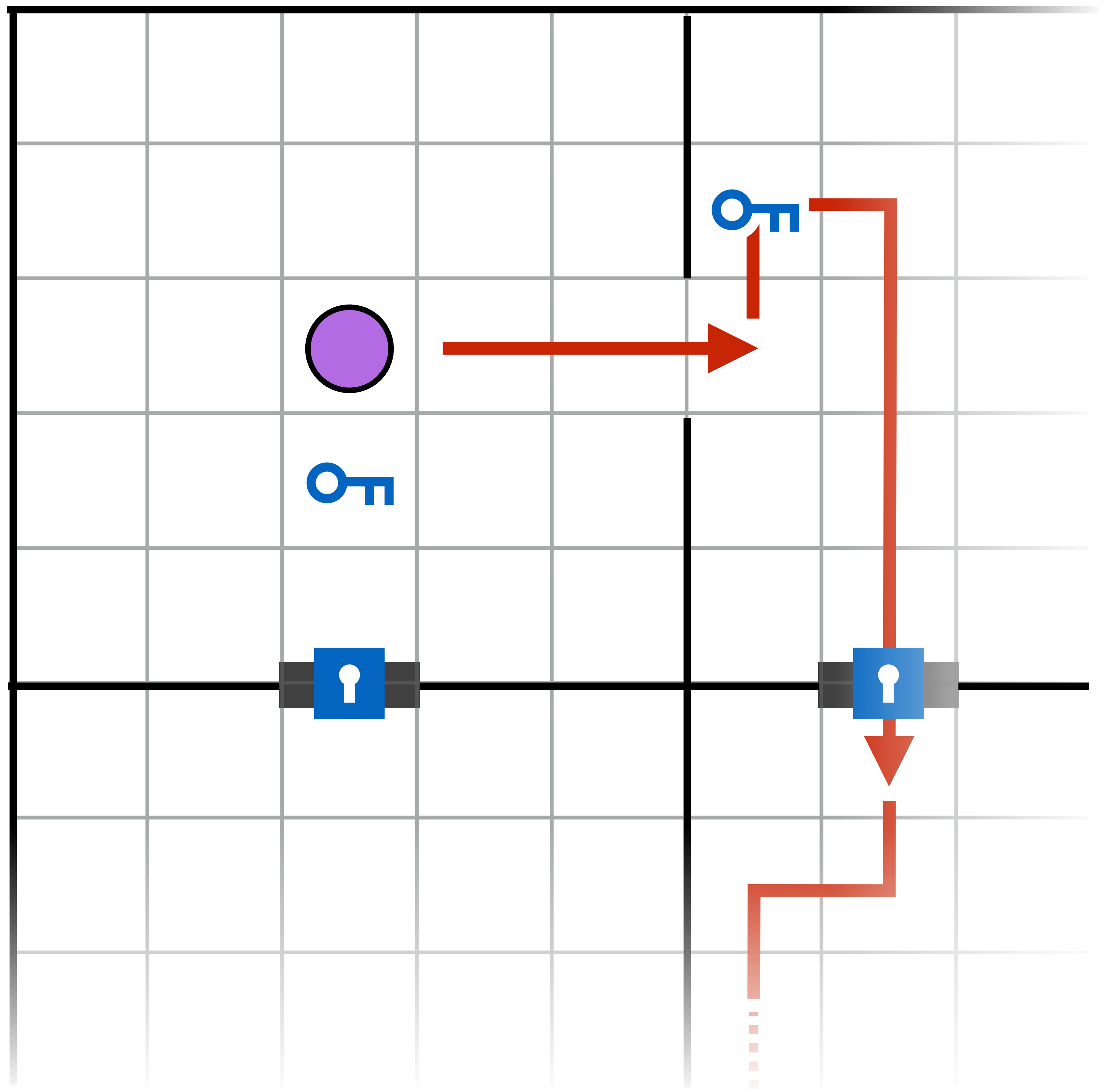
$$\sum_{\text{tasks}} \; \sum_{\text{steps}} \left(\nabla \log \boxed{\text{SUBPOLICY}}\right) \left(r_t - \boxed{\text{TASK}}\right)$$

Do sketches help?

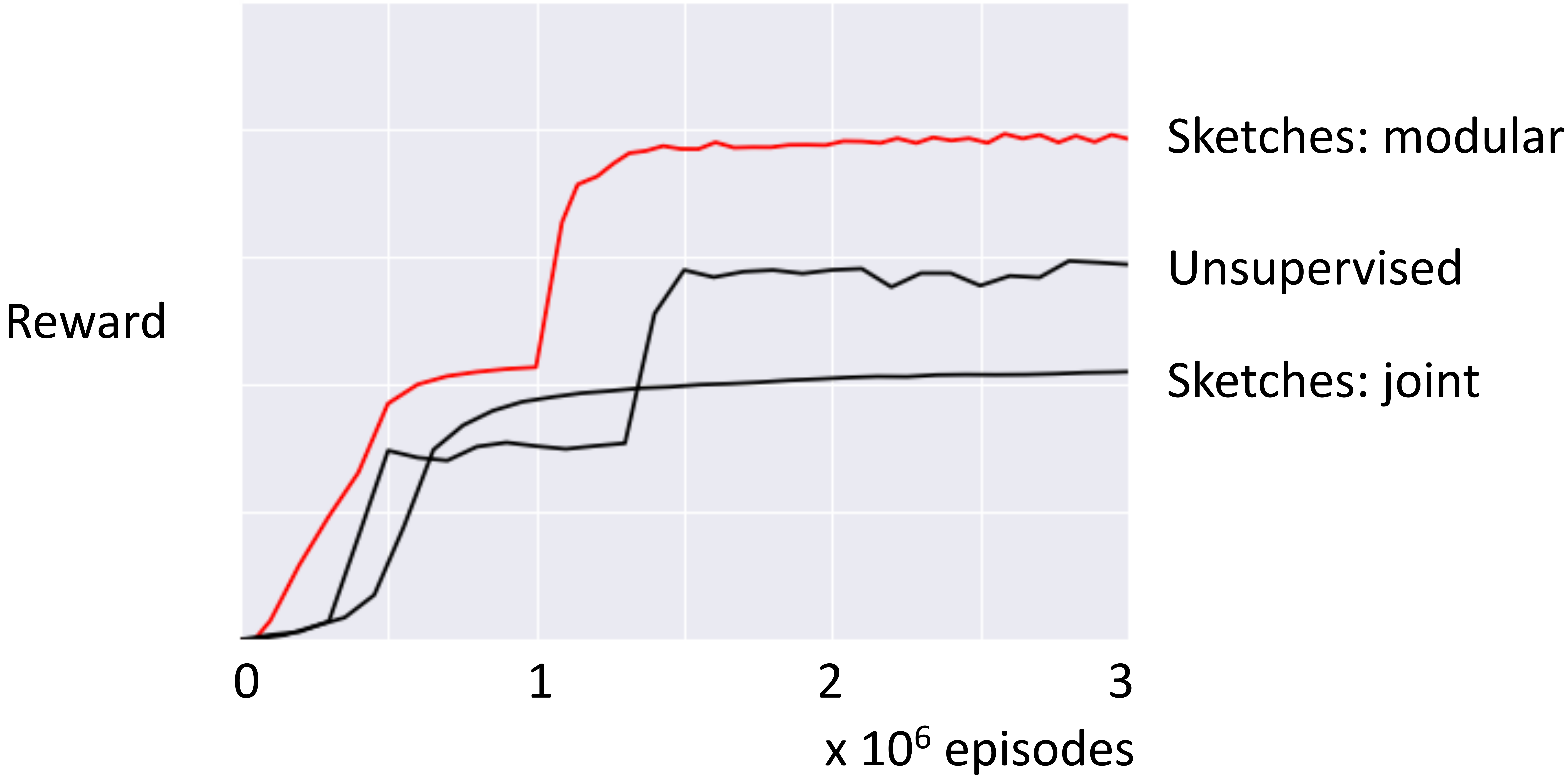
The maze navigation task



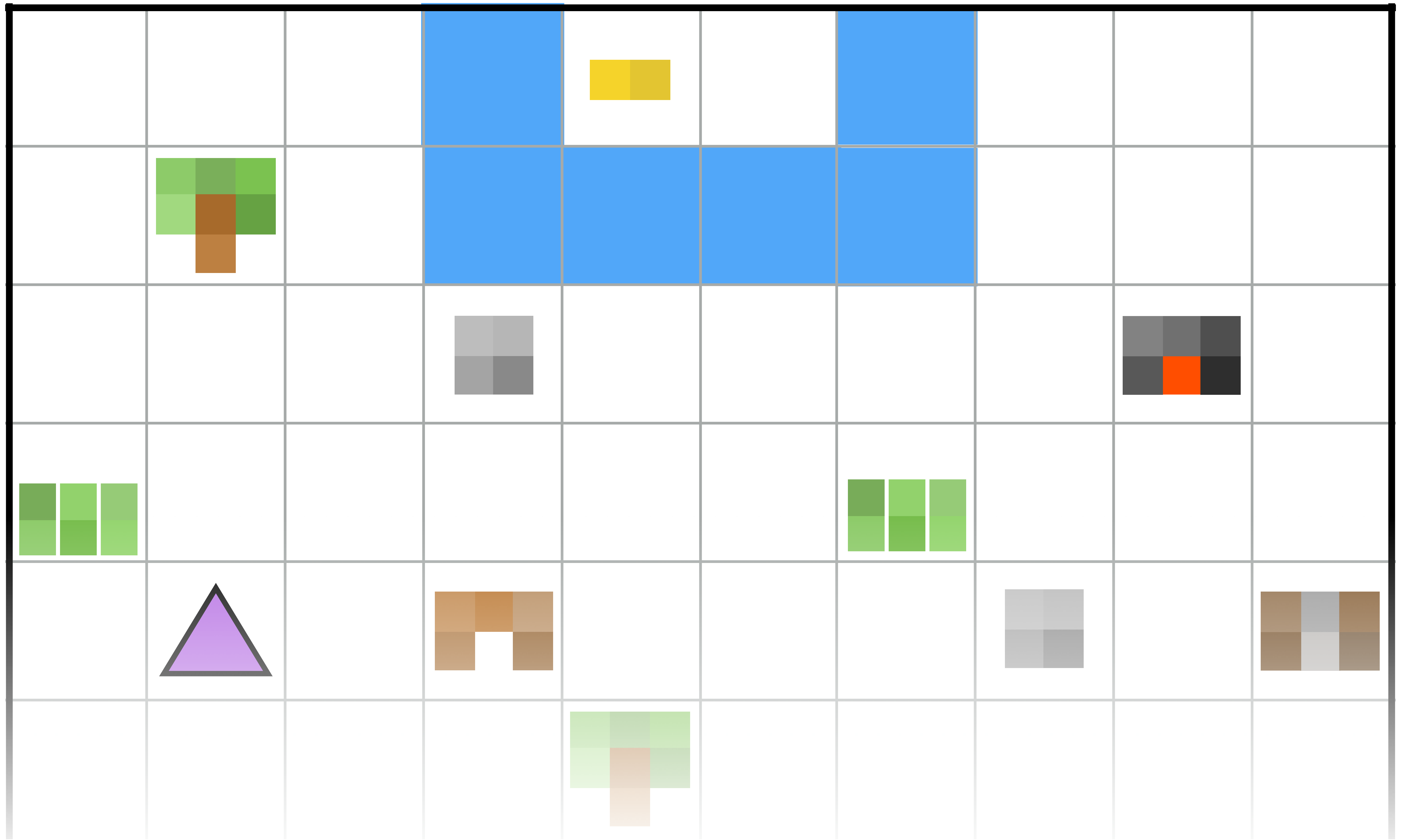
The maze navigation task



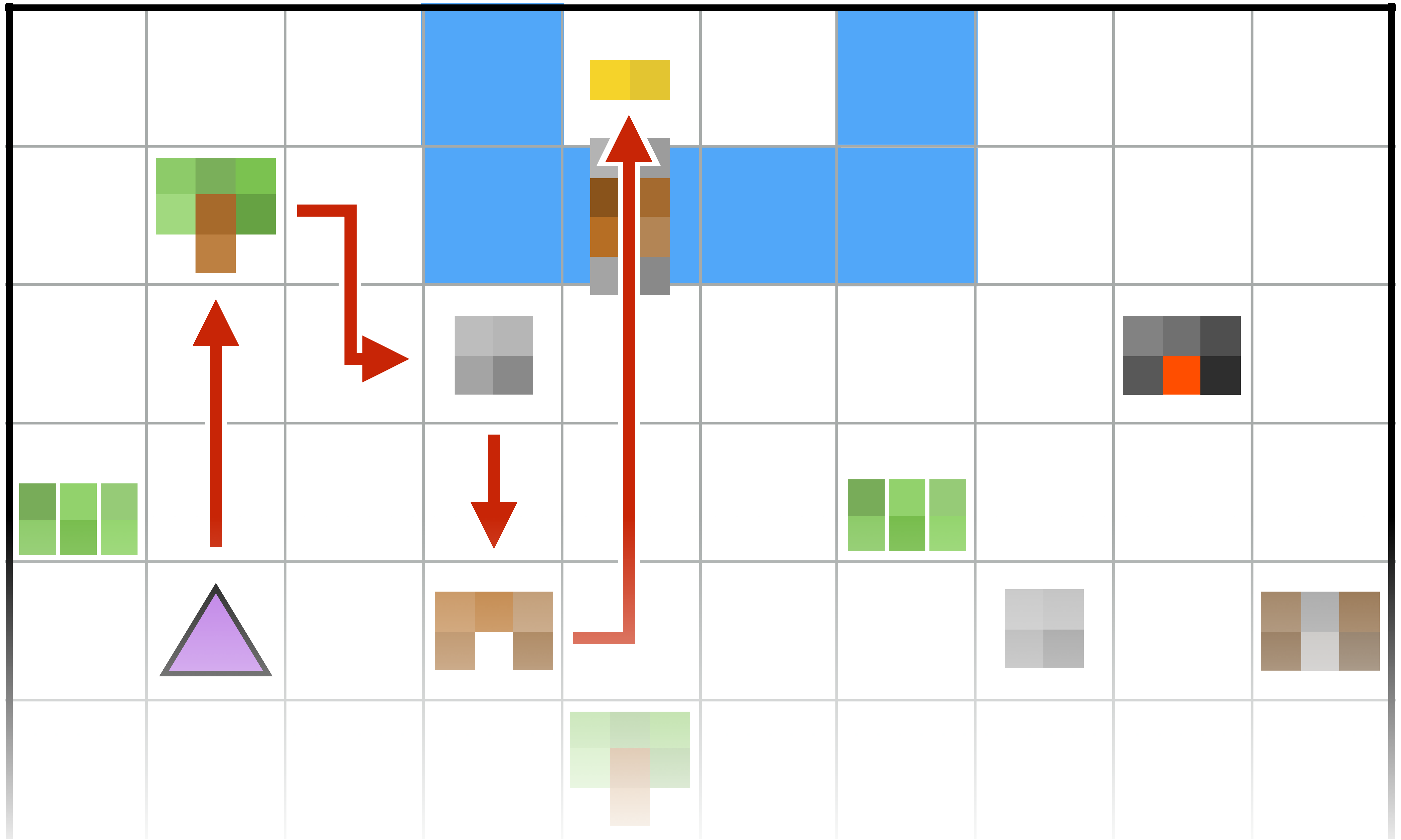
The maze navigation task



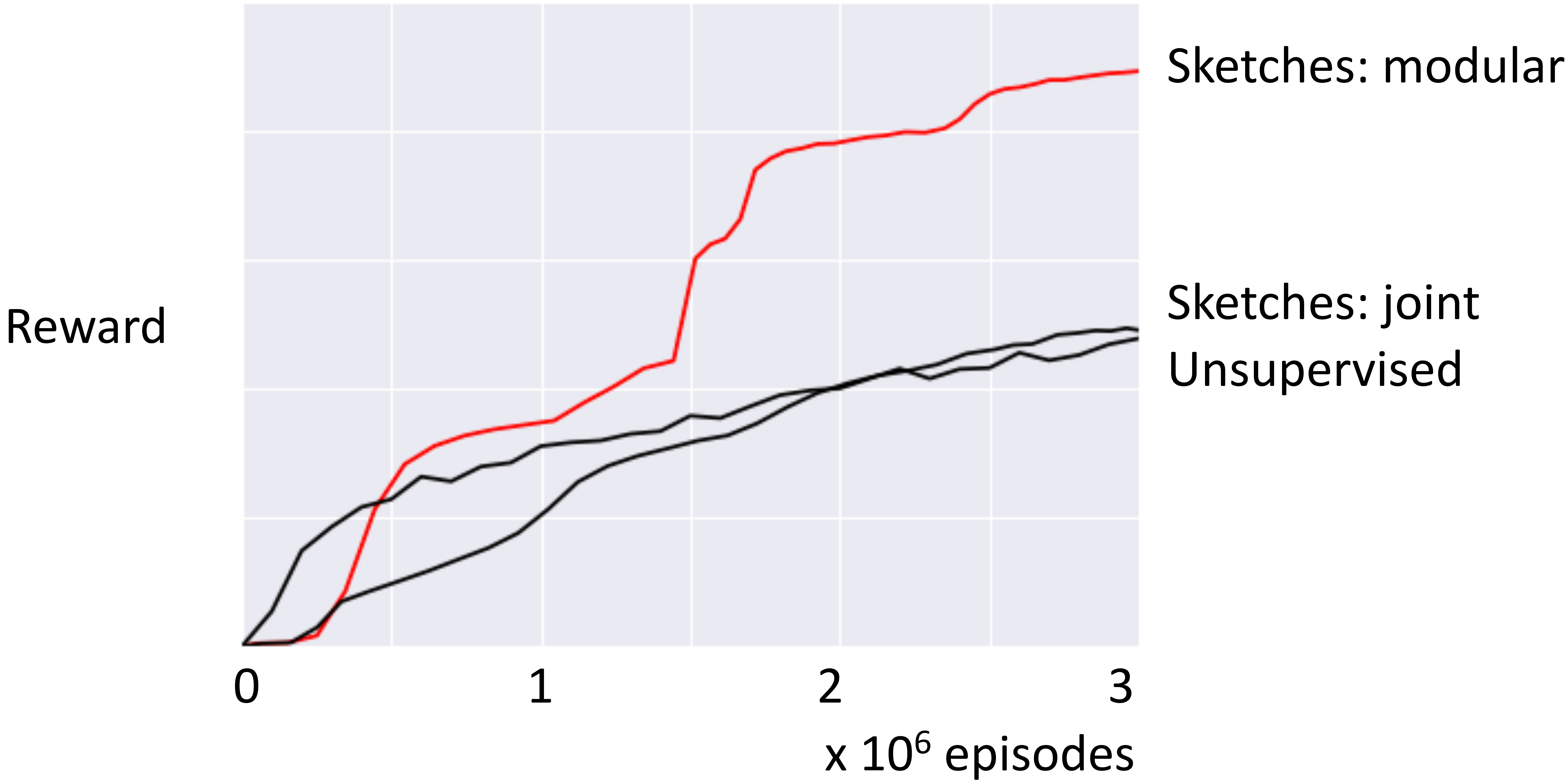
The mini-craft task



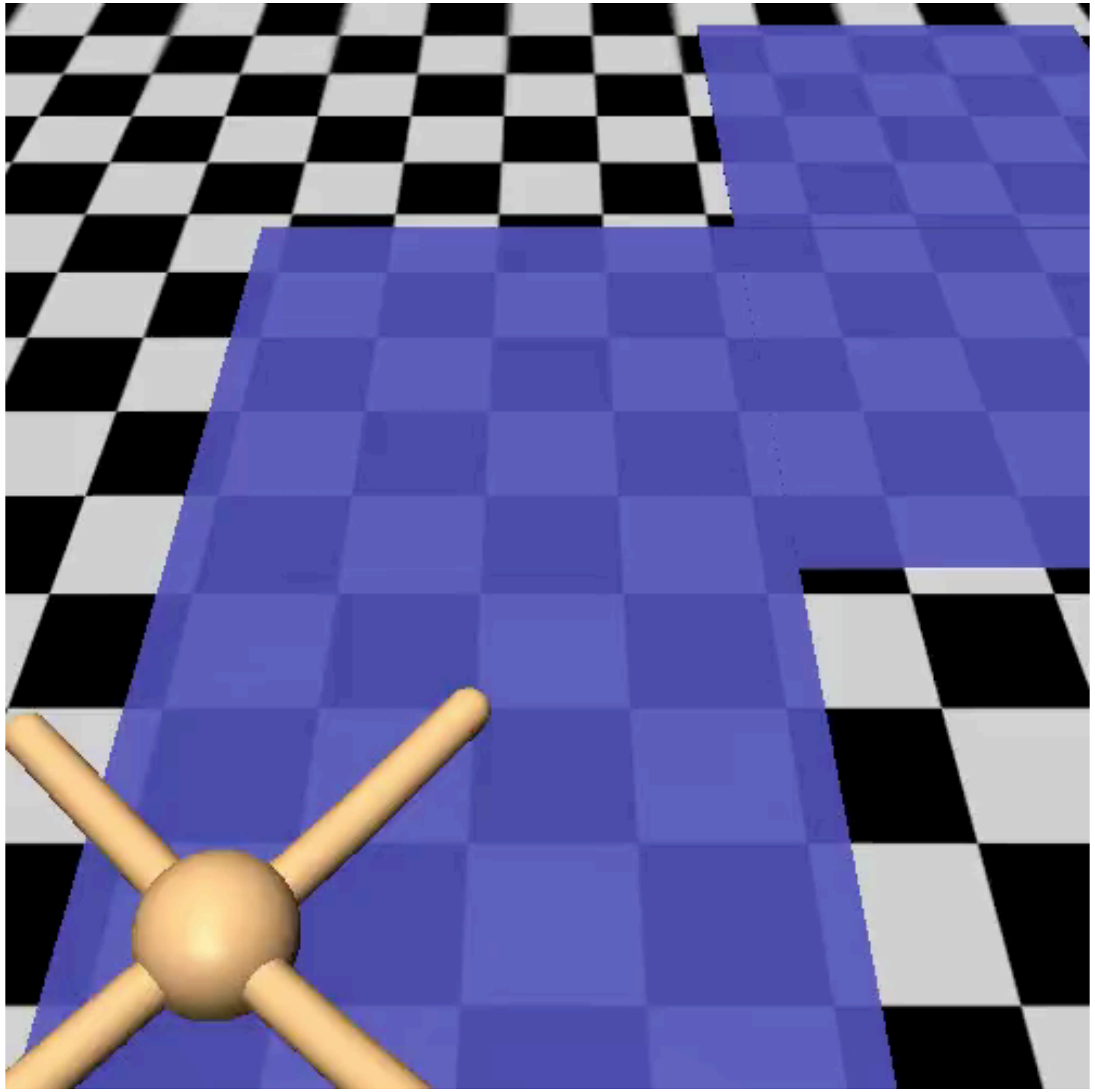
The mini-craft task



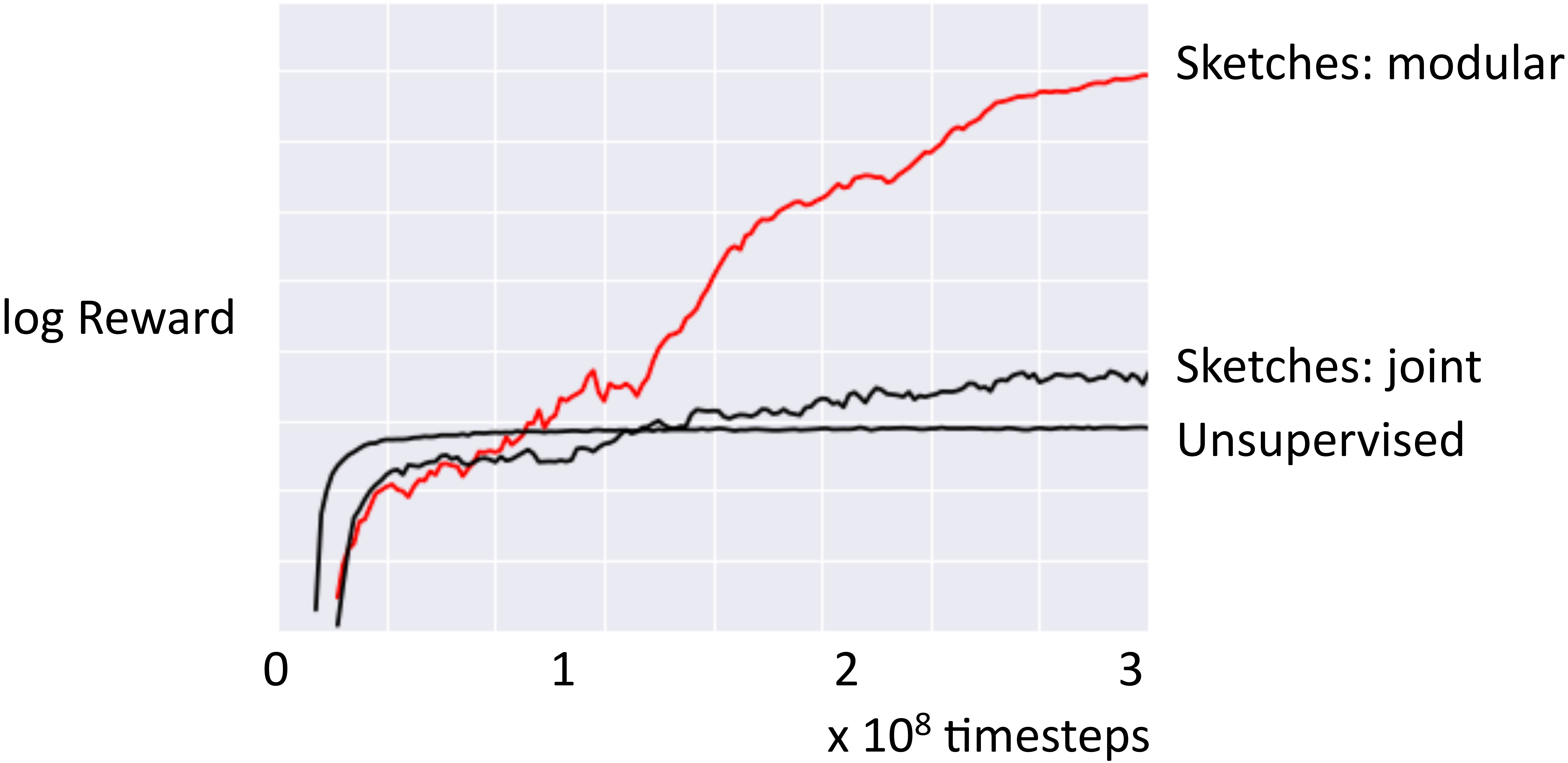
The mini-craft task



The cliff-walking task



The cliff-walking task



Zero-shot generalization

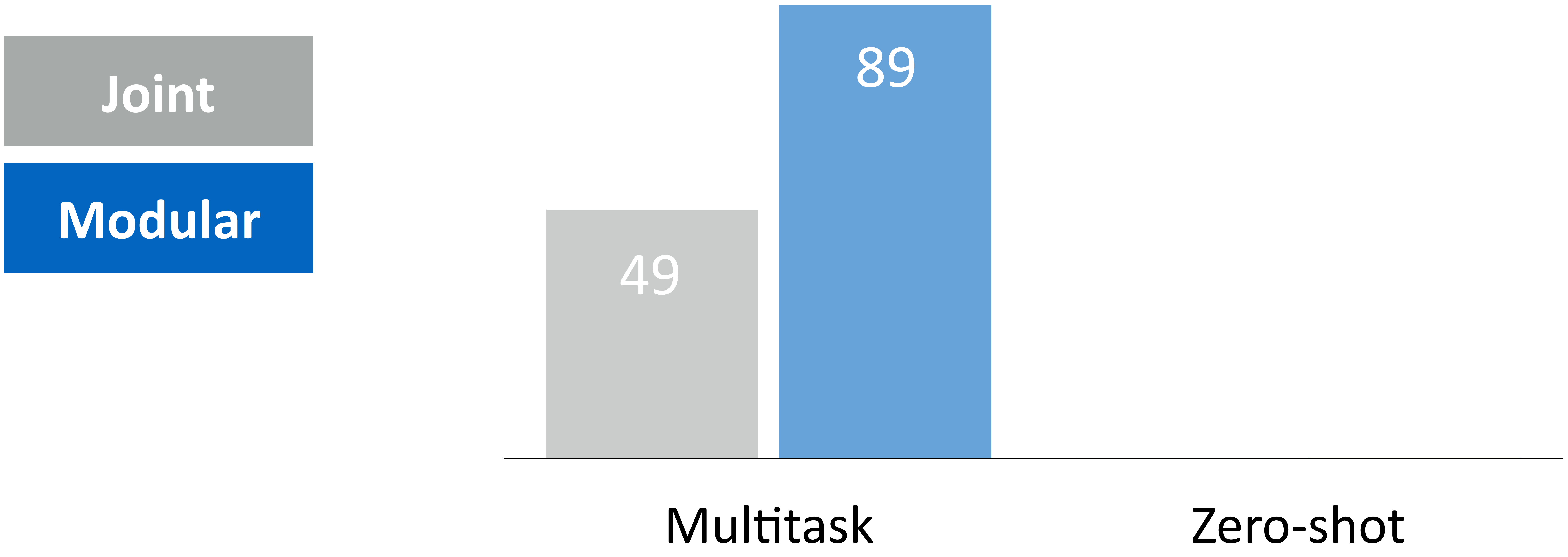
What if I see a sketch I've never seen before?

get iron

use axe

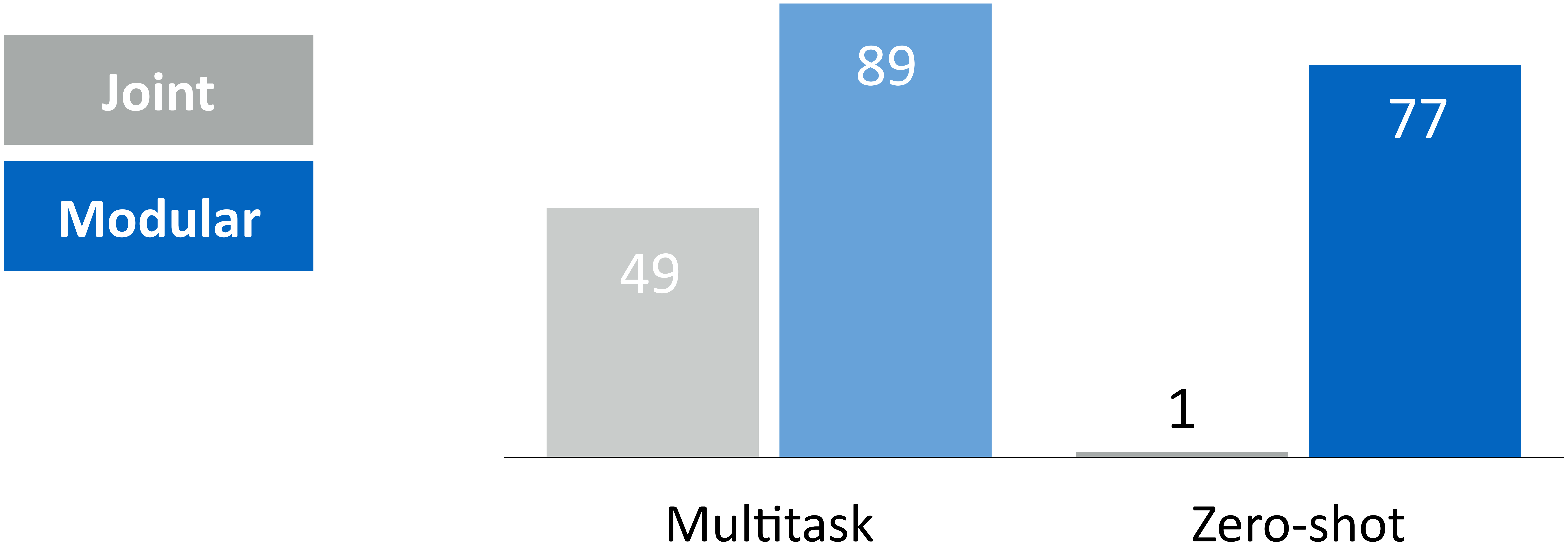
Zero-shot generalization

What if I see a sketch I’ve never seen before?



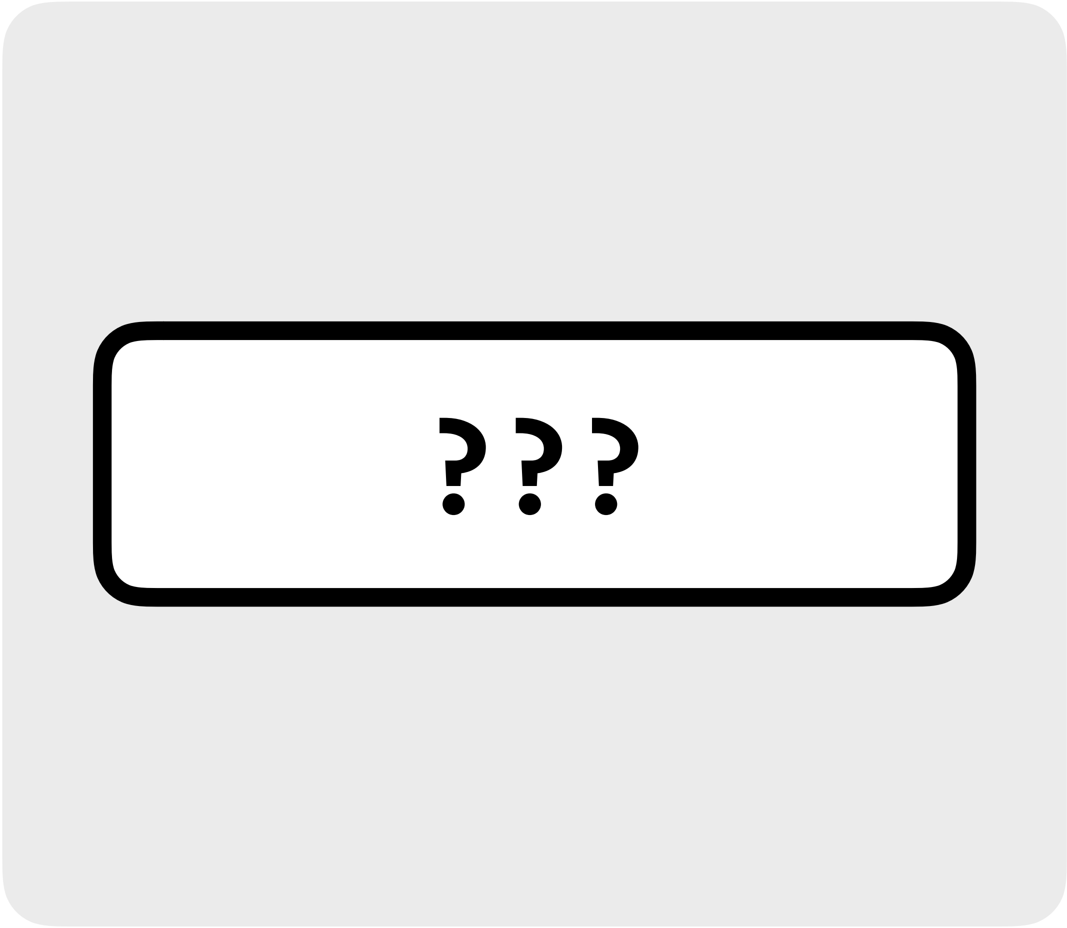
Zero-shot generalization

What if I see a sketch I’ve never seen before?



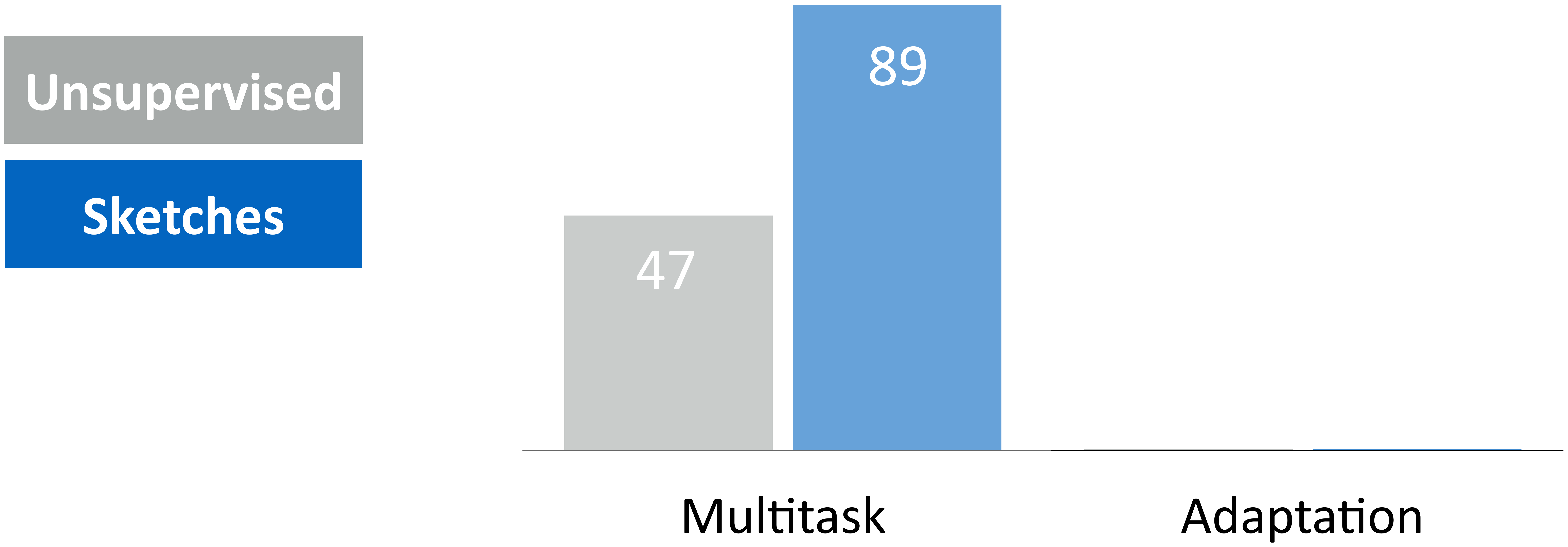
Fast adaptation

What if I don't get a sketch at test time?



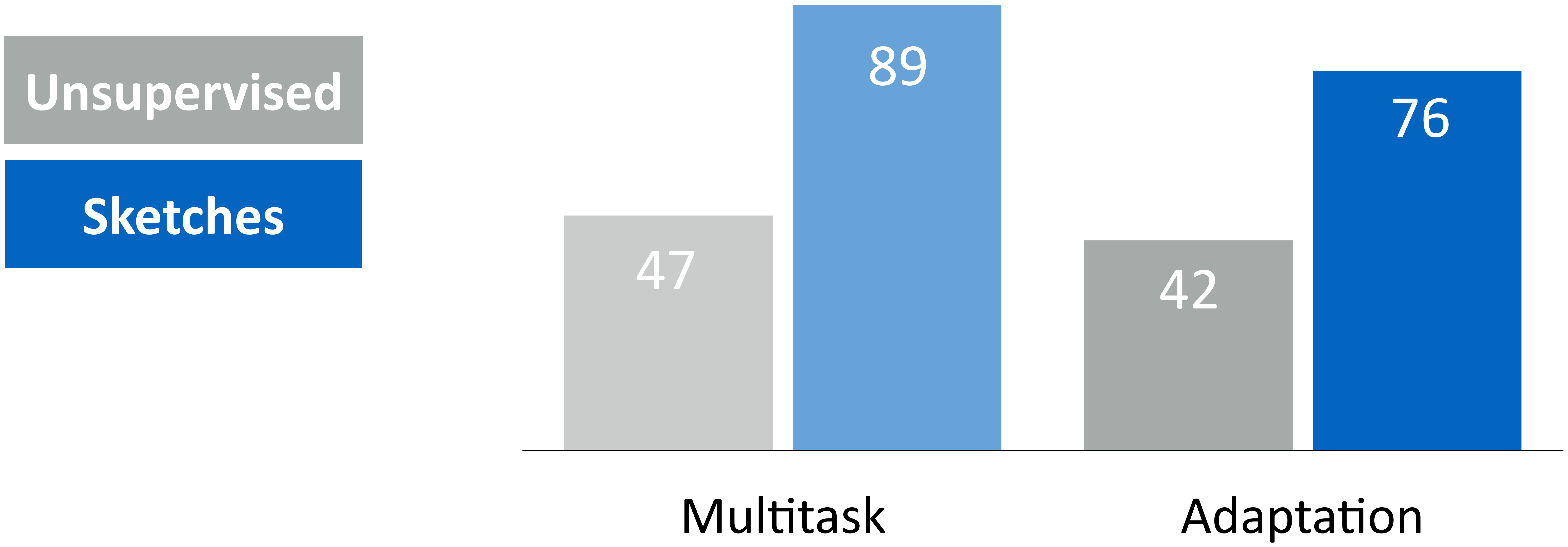
Fast adaptation

What if I don’t get a sketch at test time?



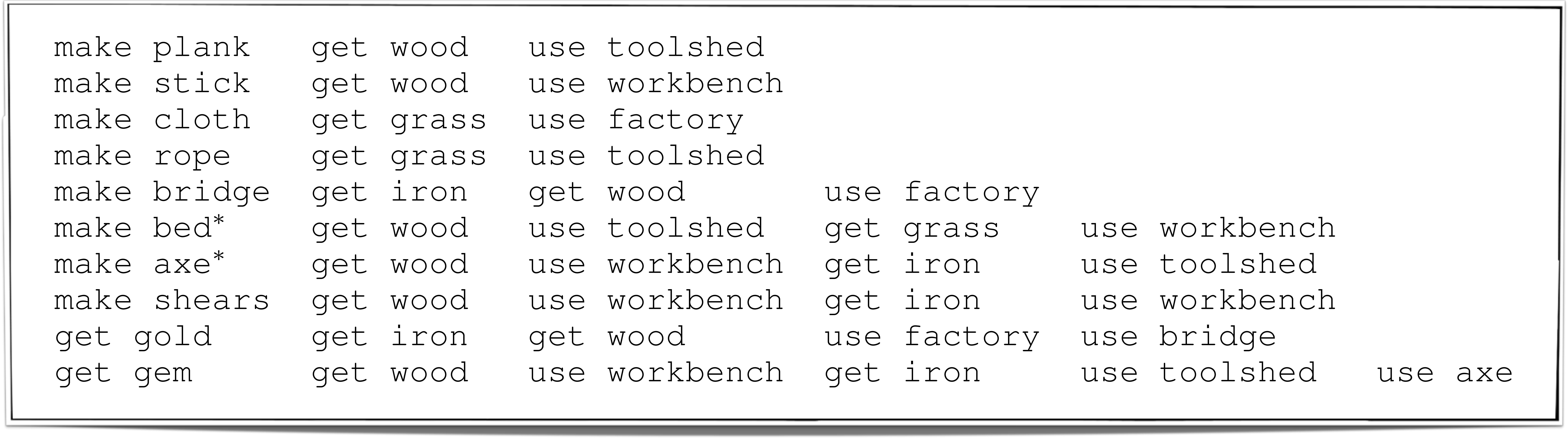
Fast adaptation

What if I don’t get a sketch at test time?



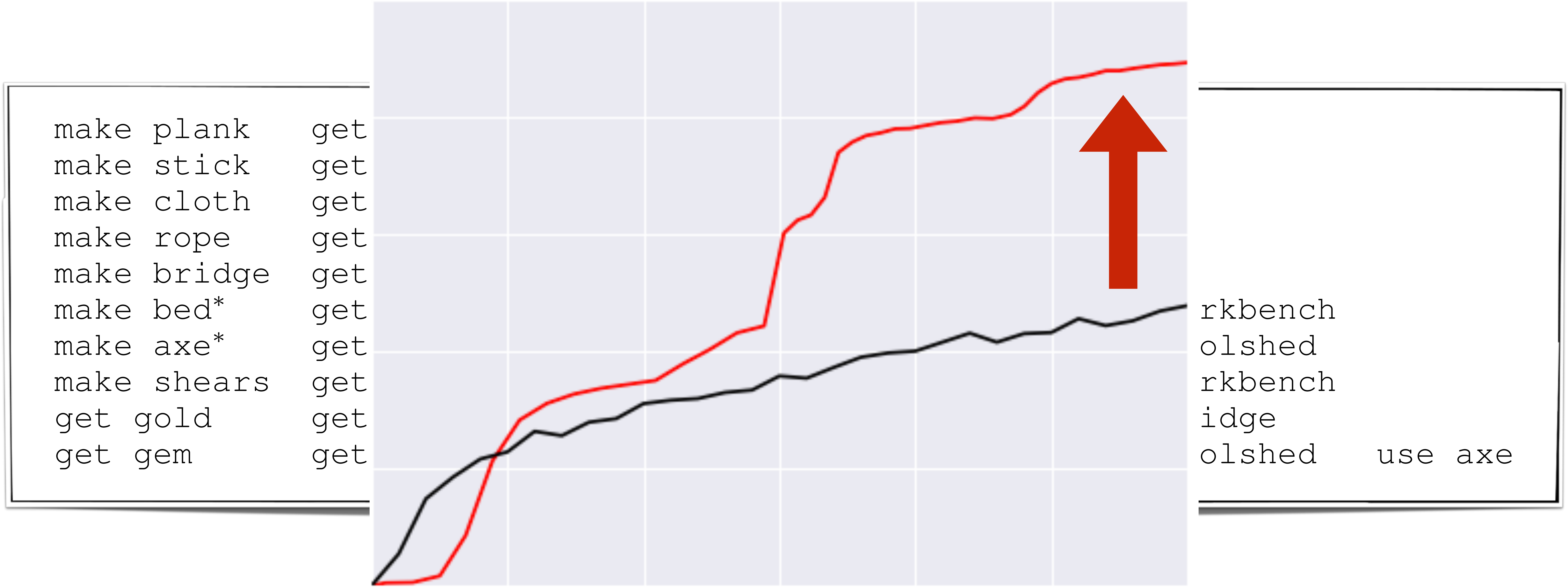
Conclusions

A tiny bit of data goes a long way



```
make plank   get wood   use toolshed
make stick   get wood   use workbench
make cloth   get grass  use factory
make rope    get grass  use toolshed
make bridge  get iron   get wood       use factory
make bed*    get wood   use toolshed   get grass    use workbench
make axe*    get wood   use workbench  get iron     use toolshed
make shears  get wood   use workbench  get iron     use workbench
get gold     get iron   get wood       use factory  use bridge
get gem      get wood   use workbench  get iron     use toolshed   use axe
```

A tiny bit of data goes a long way



```
make plank    get
make stick    get
make cloth    get
make rope     get
make bridge   get
make bed*     get                    rkbench
make axe*     get                    olshed
make shears   get                    rkbench
get gold      get                    idge
get gem       get                    olshed    use axe
```

Thank you!

https://github.com/jacobandreas/psketch