

Teaching an Agent to Play PlayStation

Play. Learn. Measure.

Crash Bandicoot: Warped · PS1 · reinforcement learning

- Reusable PlayStation RL infrastructure · Crash Bandicoot: Warped case study
- Pixels in · button presses out · rewards from game memory
- Build the system, inspect the behavior, test the claims

5 Adventures, 5 Control Problems



Toad Village

Level 1



Under Pressure

Level 2



Orient Express

Level 3



Bone Yard

Level 4



Makin' Waves

Level 5

- Toad Village · Under Pressure · Orient Express
- Bone Yard · Makin' Waves
- Run, jump and spin · swim, ride and steer · reach the exit

Orient Express: A Completed Route



Warm Start · Completed Route

Seed 0 · greedy · level complete

46s excerpt · Source 00:02-00:22 → 01:10-01:36

Orient Express · a completed route from orient_express-warm-s0.
Selected excerpts at recorded speed · source-time labels show the edits.

- Demonstration initialization · seed 0
- Watch the jumps, obstacles, and final portal approach
- Recorded outcome: level completion

Make the Game an Experiment



Policy input: images. Reward measurement: game memory.

Checkpoint + starting state + controller + configuration = an inspectable experiment

- Pixels guide actions; game memory measures reward
- Configured rewards score game events; each level defines its controller
- Reusable emulator interface · game-specific controls, rewards, and reset logic

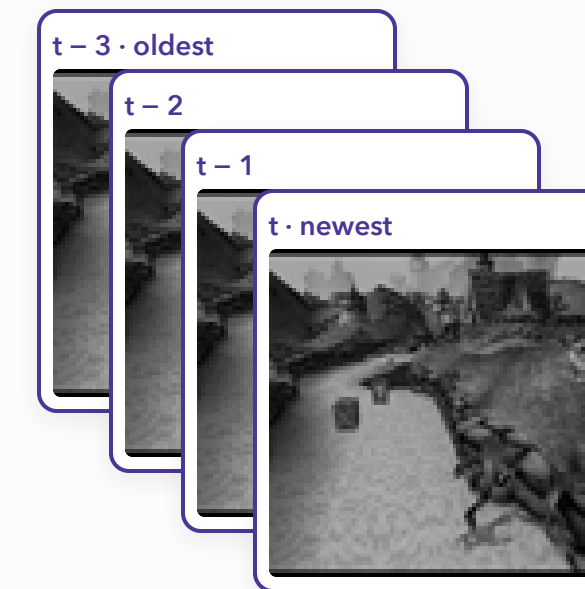
What the Network Actually Sees



Recorded color frame



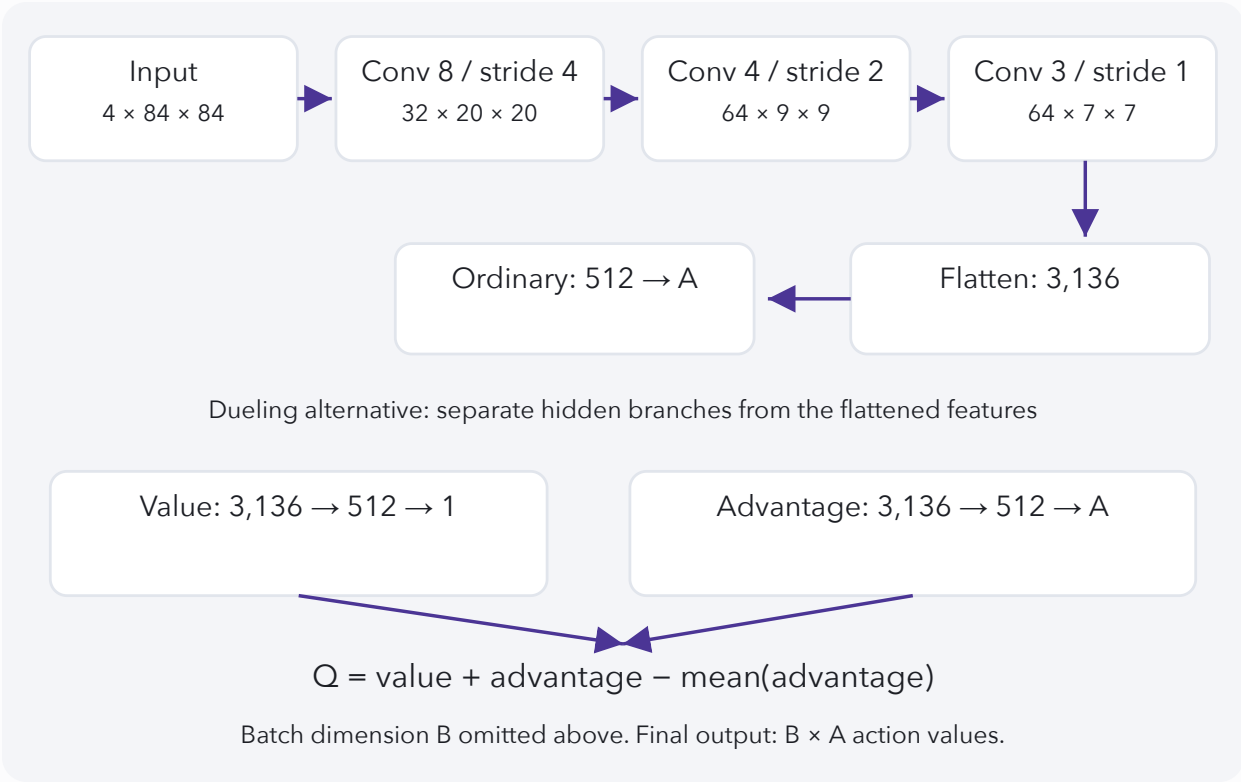
84 × 84 grayscale illustration



4 × 84 × 84 · oldest → newest
Illustration from successive video frames

- Color frame → grayscale → 84 × 84
- 4 sampled observations provide short motion history
- Illustration below uses a frame from recorded gameplay

Inside the Visual Network

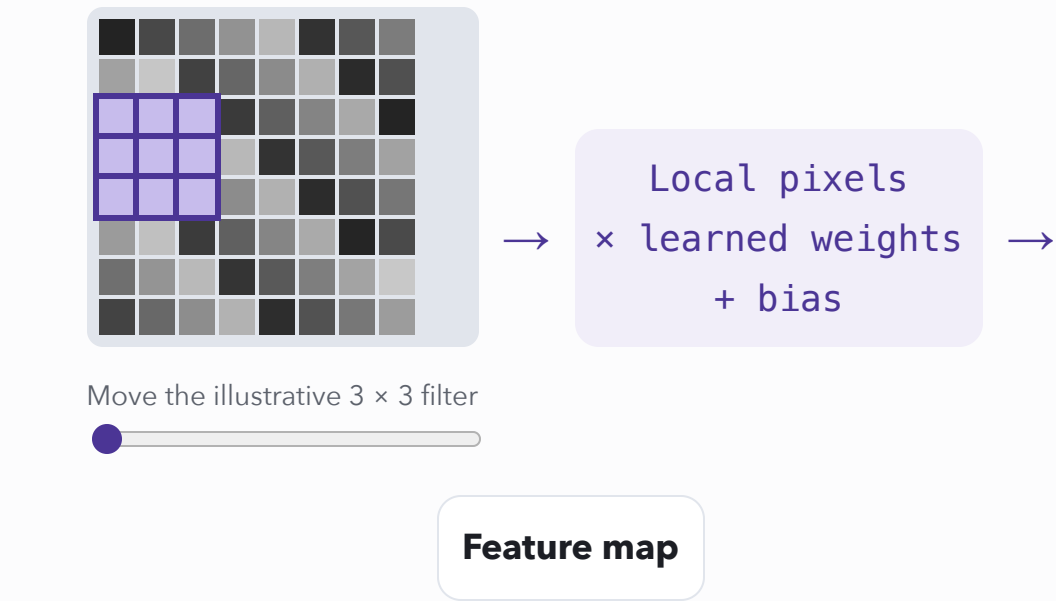


Network shapes: observations → learned features → action values

CNN backbone adapted from [Mnih et al. \(2015\)](#) · output heads adapted for PS-Env

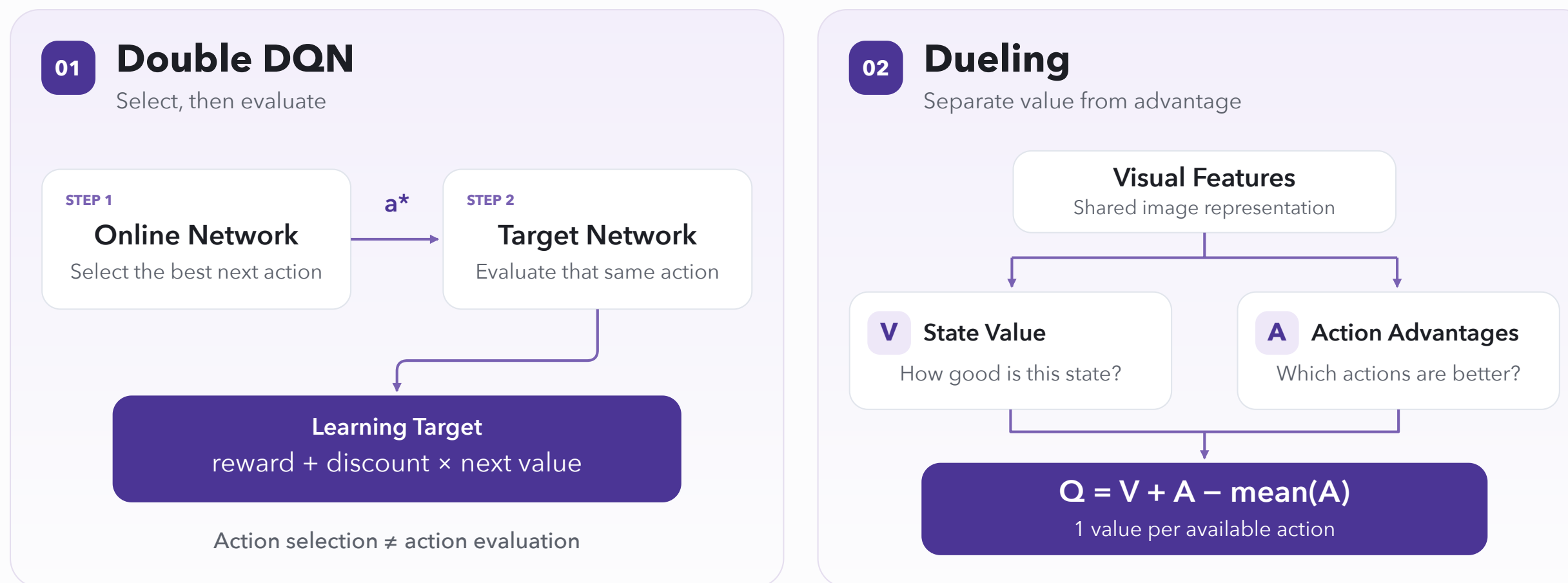
- CNN backbone adapted from Mnih et al. (2015), including David Silver
- Deeper layers combine features across larger image regions
- The output assigns a value to each available action

A Convolution Combines Local Evidence



Schematic illustration, not measured learned filters or activations

Double DQN and Dueling



Independent choices: how the target is computed, and how action values are represented.

- Our gameplay trials use both: Double DQN targets + dueling action-value heads
- Double DQN: select with the online network, evaluate with the target network
- Dueling: combine state value and action advantages into 1 Q-value per action

Reward Needs Inspection

Return

How much reward accumulated?

Coverage

Where did the policy go?

Completion

Did it reach the verified ending?

- Game events → reward components
- Return, route coverage and completion answer different questions
- Watch behavior alongside the numbers

Experiments

Training Methods

From Scratch

Learn from the agent’s own experience.

Warm Start

Learn recorded actions first; then use own experience.

DQN With Demonstrations

Warm start + examples and imitation loss during training.

Reverse Curriculum

Same learner + practice later sections first, then earlier.

Shared Policy

Setup

1 CNN trained across levels, with a separate action head for each.

Training

DQN with demonstrations; balanced samples from each level.

Question

What can the same model do on each level?

GPU Scaling

Change

1, 2, or 4 active GPUs.

Hold Fixed

Global decisions, learning updates, and batch size.

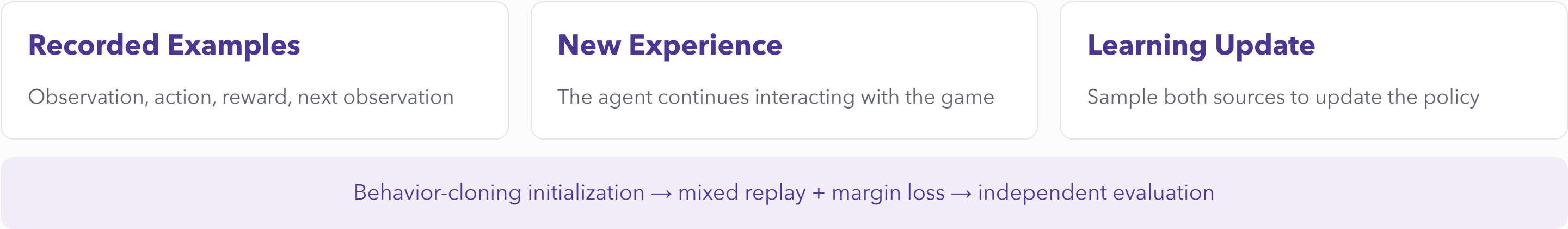
Measure

Training time for the same workload.

Gameplay is evaluated separately on each level: completion, progress, and failure mode.

- Training methods change how demonstrations and practice starts are used
- Shared-policy training combines experience across levels
- GPU scaling measures time at a fixed learning workload

DQN With Demonstrations



Adapted from DQfD · [Hester et al. \(2018\)](#) · PS-Env uses behavior-cloning pretraining and a fixed replay mixture.

- Demonstrations generated by automated beam search in the emulator
- Mix recorded transitions with the agent’s own experience
- Reward-based loss + imitation loss on demonstration samples

Reverse Curriculum Learning

Practice a Section

Restore a saved state later in the route

Move Earlier

Success moves the practice frontier toward the start

Test the Full Route

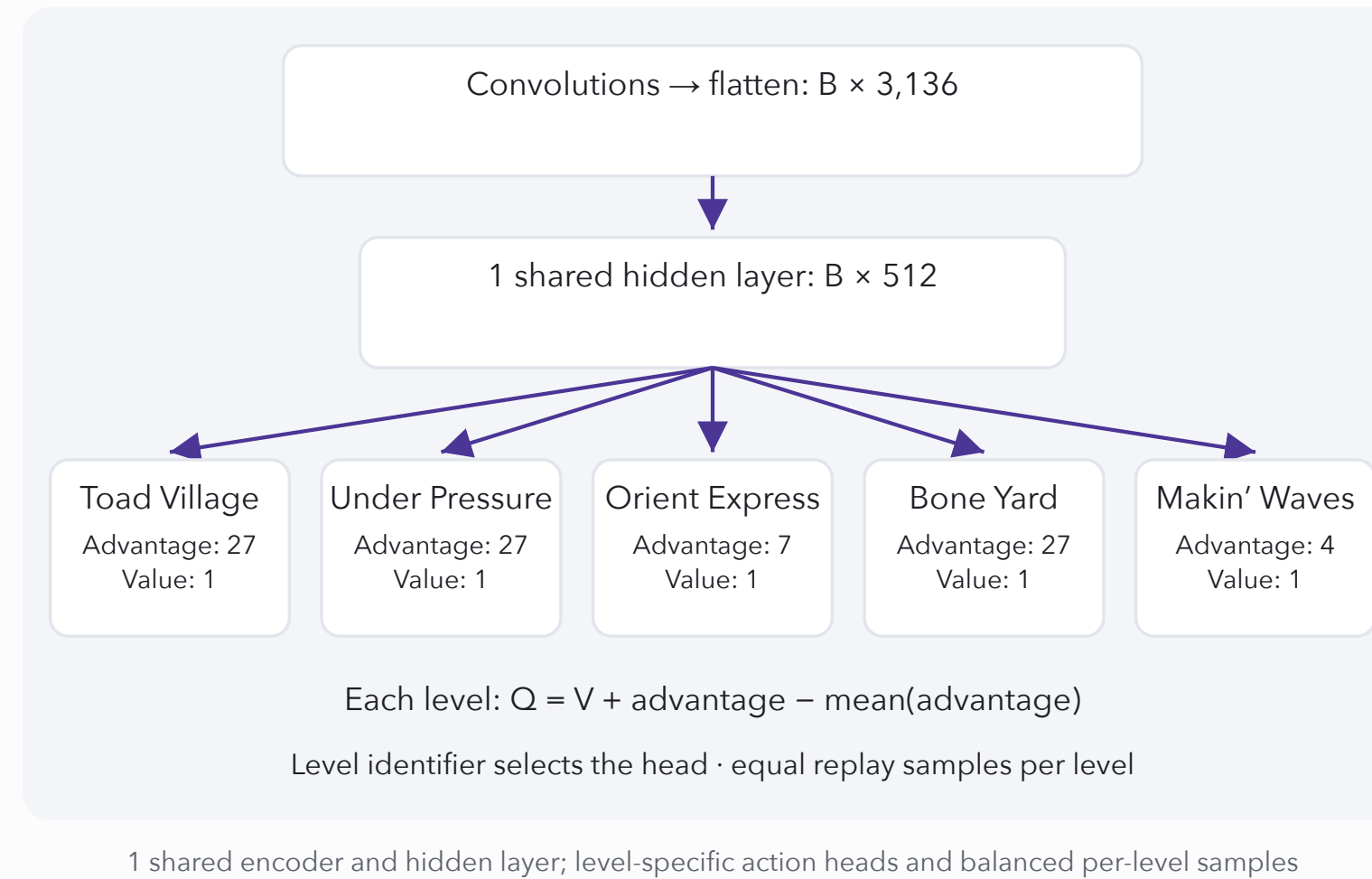
Evaluate completion from the original start

This adaptation: behavior cloning + DQN learner with demonstrations + reverse curriculum

Inspired by [Salimans and Chen \(2018\)](#) · adapted to local section goals and a DQN learner with demonstrations.

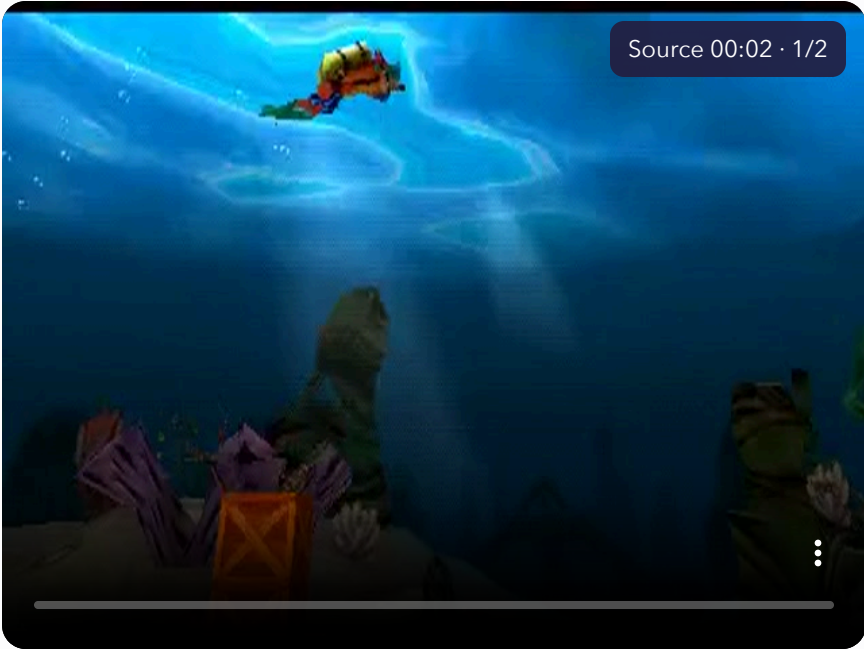
- Demonstration-state curriculum · inspired by Salimans and Chen (2018)
- Practice later sections first, then move the starting frontier earlier
- PS-Env adaptation: local section goals plus full-start episodes

Shared Architecture Across Adventures



- 1 shared CNN and 512-unit representation
- Level identifier selects a separate action-value head
- Balance training samples; test the trained model on each level

Under Pressure: Progress Without Completion



Scratch

Seed 0 · greedy · death

32s excerpt · Source 00:02-00:14 → 01:22-01:42



Reverse Curriculum Learning

Seed 0 · greedy · decision limit

30s excerpt · Source 00:14-00:32 → 09:48-10:00



DQN With Demonstrations

Seed 0 · greedy · decision limit

30s excerpt · Source 00:14-00:32 → 09:48-10:00

Under Pressure · an underwater adventure. Reaching the reference-route milestones does not establish level completion.
Selected excerpts at recorded speed · source-time labels show the edits.

- Scratch: movement around rocks and crates; ends in death
- Curriculum and demonstrations: similar machinery-and-cavern route
- Both reach all 8 reference milestones, but not level completion

Orient Express: Progress Versus Completion



Orient Express • Scratch

Seed 0 · greedy · death

32s excerpt · Source 00:03-00:15 → 01:43-02:03



Orient Express • Reverse Curriculum

Seed 0 · greedy · level complete

34s excerpt · Source 00:08-00:20 → 01:14-01:36

Same recorded trajectory as slide 3

Scratch collects new experience; warm start initializes weights; DQN with demonstrations reuses demonstrations; reverse curriculum changes practice starts.
Selected excerpts at recorded speed · source-time labels show the edits.

- Scratch: moves along the route, but the evaluation ends in death
- Curriculum export: reaches the final portal; completion logged
- Right-hand video repeats the trajectory from slide 3

1 Shared Policy, Different Outcomes



Shared Policy · Orient Express

Seed 1 · greedy · level complete

30s excerpt · Source 00:12-00:24 → 01:18-01:36

Same recorded trajectory as slide 3



Shared Policy · Under Pressure

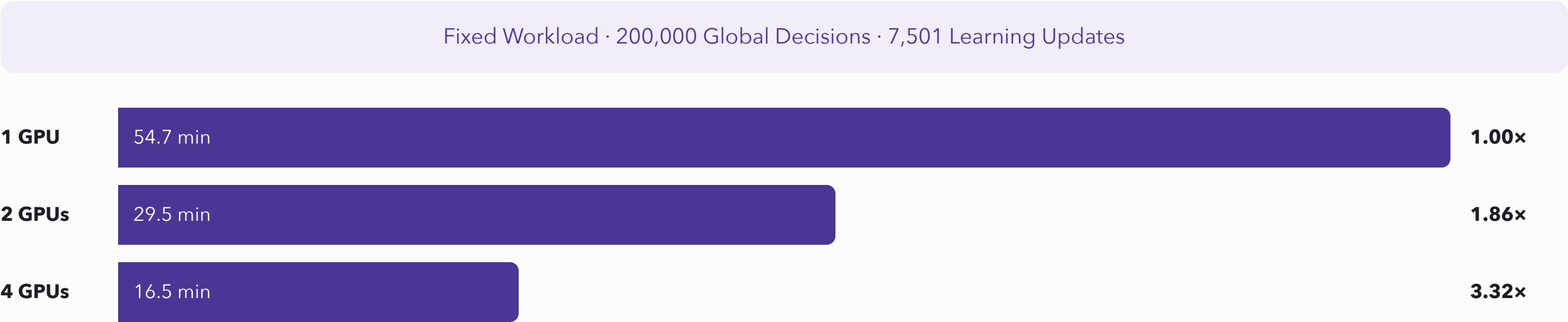
Same experiment · greedy · decision limit

28s excerpt · Source 00:14-00:30 → 09:48-10:00

1 shared encoder, level-specific action heads. 1 level completing does not imply 5-level mastery.
Selected excerpts at recorded speed · source-time labels show the edits.

- Orient Express: completion logged; repeated trajectory from slide 3
- Under Pressure: movement continues, but the decision budget runs out
- Evaluate the shared model separately on each level

Same Workload, Faster Training

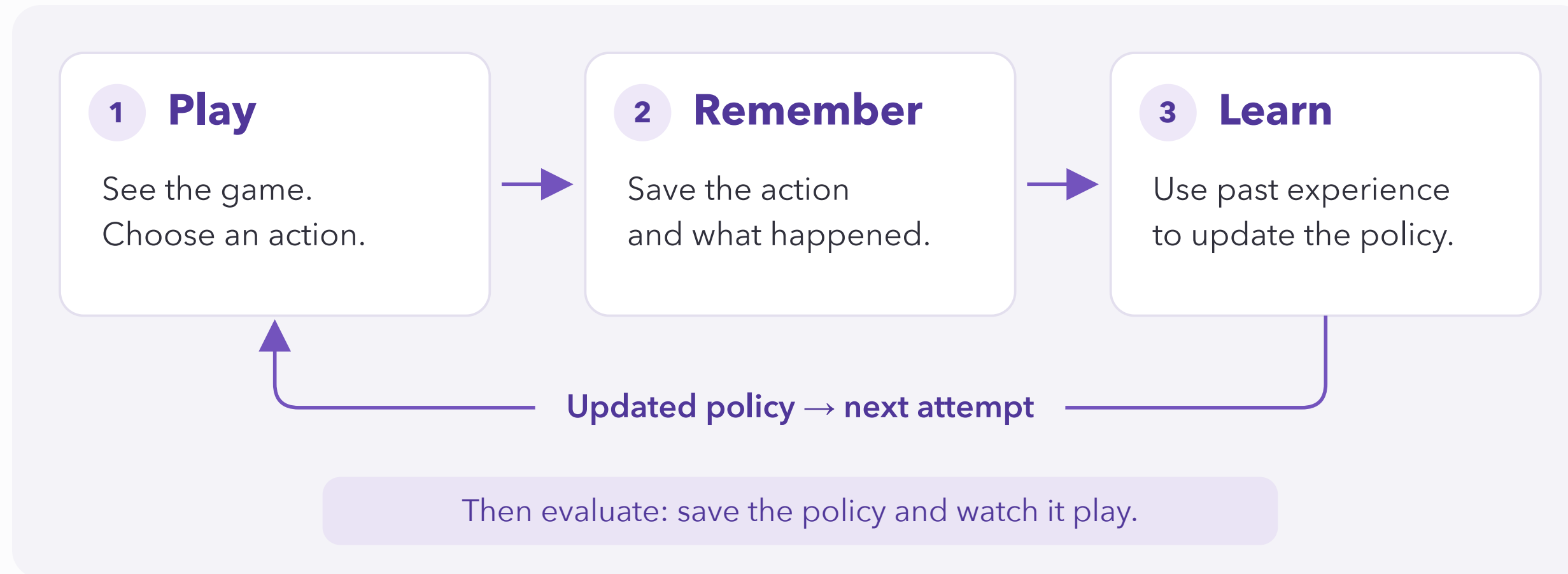


Mean recorded training interval · 2 seeds · provisioning and separate evaluation excluded

3.32x training speedup · same workload

- Same work: 200,000 global decisions and 7,501 learning updates
- 4 GPUs: 54.7 → 16.5 minutes, about 3.32x faster
- Training time only · same host · 2 seeds per GPU setting

How Playing Becomes Learning



Play, remember, learn—and test the saved policy.

- Images → action values → button presses
- Experience → training updates → a changed policy
- Save the policy, then test what it learned

A Childhood Game, a New Starting Point

Working Platform

Real game interaction, saved policies, and inspectable runs.

Faster Training

3.32× speedup on 4 GPUs for the tested workload.

A Route Completed

A separate demonstration-trained policy finished Orient Express.

A working foundation for learning to play further—and measuring every step.

- Built a working platform for repeatable, inspectable gameplay experiments
- Measured 3.32× faster training and recorded an Orient Express completion
- Next: turn this foundation into more reliable play across adventures