A AGENT AND ENVIRONMENT DETAILS

A.1 ATARI

We used the OpenAI Baselines implementation of DQN with default settings with all Atari environments. The only modification to the model was the concatenation of any state augmentation (where employed) with the output of the convolutional layers. From a previous set of experiments, we found that the use of prioritized replay dramatically shrinks the optimal values of the reward shaping coefficient, but that the optimal choice is still strongly based on the choice of environment. We hypothesize that this is largely due to the differing MDP dynamics and reward frequency.

A.2 MUJOCO

We used the unmodified OpenAI baselines implementation of PPO in all MuJoCo environments, simply appending any state augmentation to the MDP observation as input to the model.

B CONSTRAINTS AND TRANSLATION FUNCTIONS

B.1 ATARI BREAKOUT PADDLE BALL DISTANCE

We designed the "paddle ball distance" constraint for the Breakout environment. The translation function uses the last frame from each environment observation and calculates, using rudimentary computer vision provided by scikit-image van der Walt et al. (2014), the horizontal distance between the center of the ball and the center of the paddle. Then, if the paddle was too far to the left relative to the ball and the "move right" action wasn't taken that frame, the "L" token is output and similar for the "R" token. Otherwise, if the ball is outside the area under the bricks or no other token is output, a zero token is output.

The constraint is a simple counter that increments an L counter for each successive L token and similar for R and accepts on 3 successive, identical tokens. Receiving a zero token or a token from the other direction resets each counter.

We chose 10 pixels as the maximum allowed distance between ball and paddle under this constraint, but this turned out to be too restrictive for good performance under hard constraints.

B.2 ATARI: SPACE INVADERS

The translation function for "dangerzone" uses scikit-learn to find the position of each downward-travelling bullet fired by enemies in the area immediately above the player's ship, as well as the position of the player's ship. Then, it calculates the distance between the closest bullet and the ship, as well as whether the bullet is to the right, left, or above the ship. Each token is a concatenation of the direction of the bullet (left, right, or above) and a discretization of the distance ($x \leq 12$ pixels, $12 < x \leq 24$ pixels, or $x > 24$ pixels).

The constraint is violated if a bullet is at the closest distance, < 12 pixels, and an action to dodge isn't taken, i.e., moving right from a bullet to the left, moving left from a bullet on the right, or moving in either direction from a bullet above.

B.3 MUJOCO: REACHER

Two related constraints were explored in the Reacher environment: an actuation constraint similar to those presented in Atari, and a dynamic actuation constraint that was modified to take the relative position of the ball to the goal into account. The translation function each used discretizes the action in increments of 0.2. Each used a DFA to track the sum of actuation values which accepts, if the sum over 3 timesteps is greater than a threshold of 4.0. In the dynamic reacher constraint, the acceptance threshold varies by how close the goal is.

B.4 MUJOCO: HALF CHEETAH

The Half Cheetah environment uses the one dimensional no-dithering constraint (similar to Atari) on each of the six joints of the simulated robot. The translation function discretizes positive and negative joint forces to be right and left, respectively.

C SAFETY GYM REWARD SHAPING PLOTS

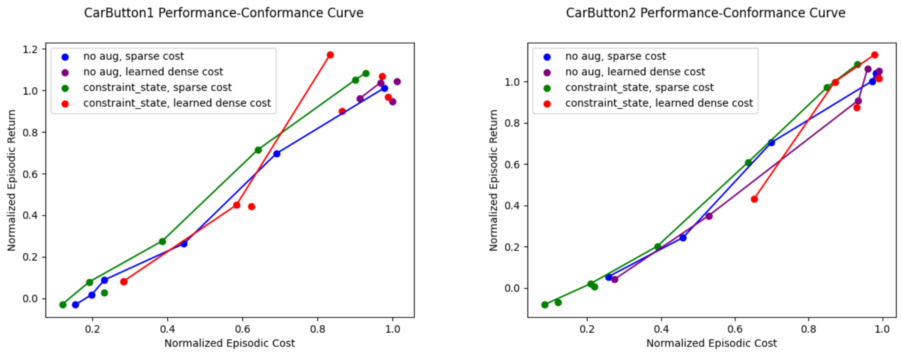

Figure 3: Performance/Conformance curves for CarButton environments, with Pareto frontiers plotter per reward shaping method.

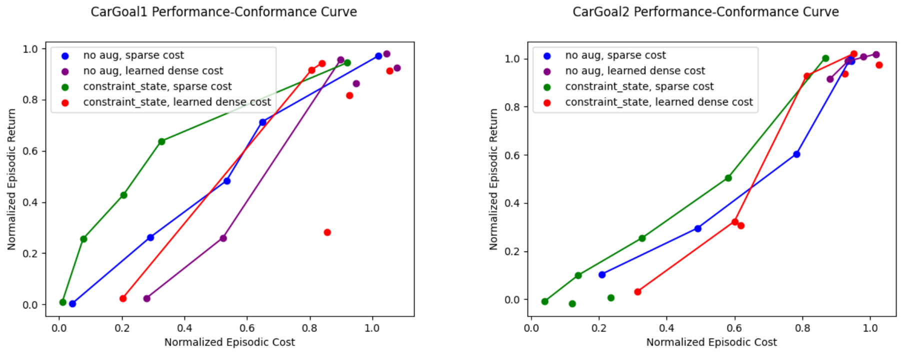

Figure 4: Performance/Conformance curves for CarGoal environments, with Pareto frontiers plotted per reward shaping method.

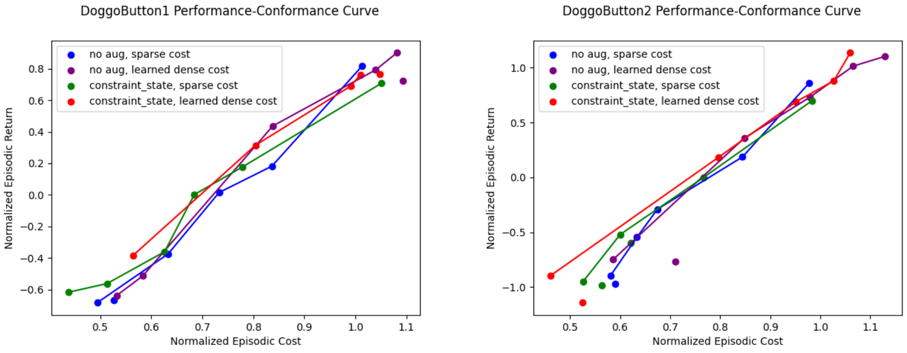

Figure 5: Performance/Conformance curves for DoggoButton environments, with Pareto frontiers plotted per reward shaping method.

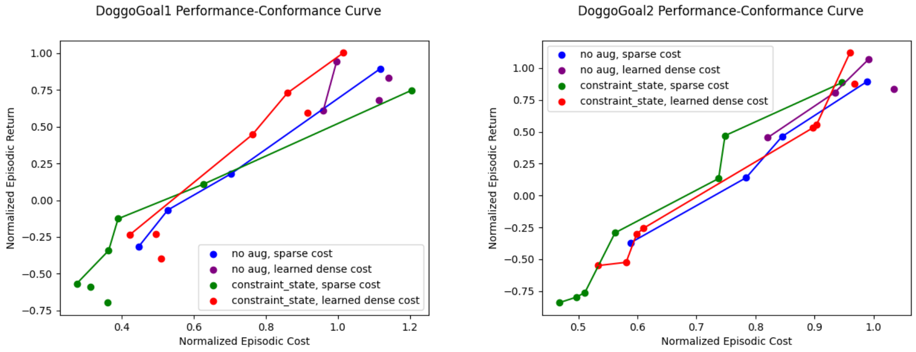

Figure 6: Performance/Conformance curves for DoggoGoal environments, with Pareto frontiers plotted per reward shaping method.

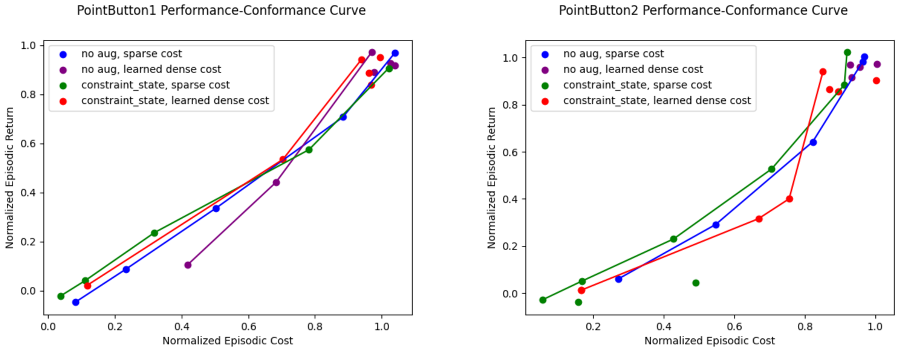

Figure 7: Performance/Conformance curves for PointButton environments, with Pareto frontiers plotted per reward shaping method.

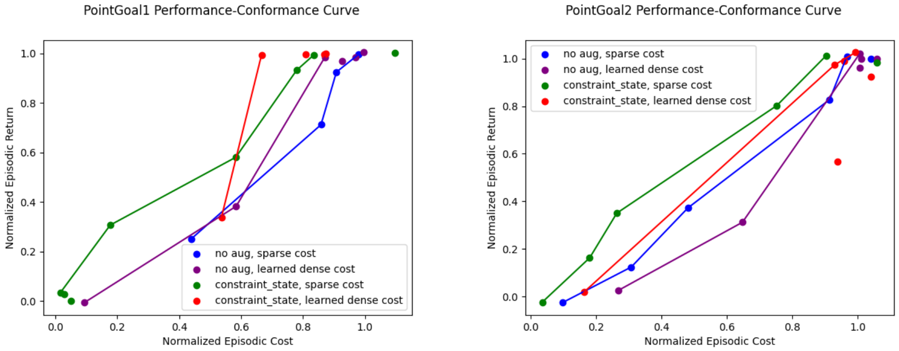

Figure 8: Performance/Conformance curves for PointGoal environments, with Pareto frontiers plotted per reward shaping method.

D FULL 2D NO-DITHERING REGEX

The no-2d-dithering constraint used with the Seaquest environment was generated by a simple Python script which generated all sequences of up to 4 moves and filtered them to those which end where they begin with no side effects (e.g., pressing the fire button).

```
((2|A) (2|A) (5|D) (5|D)) | ((2|A) (5|D) (2|A) (5|D)) | ((2|A) (5|D) (5|D) (2|A
)) | ((5|D) (2|A) (2|A) (5|D)) | ((5|D) (2|A) (5|D) (2|A)) | ((5|D) (5|D) (2|A) (
2|A)) | ((2|A) (2|A) (8|G) (9|H)) | ((2|A) (2|A) (9|H) (8|G)) | ((2|A) (8|G) (2|
A)(9|H))|((2|A)(8|G)(9|H)(2|A))|((2|A)(9|H)(2|A)(8|G))|((2|A)(9|H)
(8|G)(2|A))|((8|G)(2|A)(2|A)(9|H))|((8|G)(2|A)(9|H)(2|A))|((8|G)(9
|H)(2|A)(2|A))|((9|H)(2|A)(2|A)(8|G))|((9|H)(2|A)(8|G)(2|A))|((9|H
)(8|G)(2|A)(2|A))|((2|A)(3|B)(4|C)(5|D))|((2|A)(3|B)(5|D)(4|C))|((
2|A)(4|C)(3|B)(5|D))|((2|A)(4|C)(5|D)(3|B))|((2|A)(5|D)(3|B)(4|C))
|((2|A)(5|D)(4|C)(3|B))|((3|B)(2|A)(4|C)(5|D))|((3|B)(2|A)(5|D)(4|
C))|((3|B)(4|C)(2|A)(5|D))|((3|B)(4|C)(5|D)(2|A))|((3|B)(5|D)(2|A)
(4|C))|((3|B)(5|D)(4|C)(2|A))|((4|C)(2|A)(3|B)(5|D))|((4|C)(2|A)(5
|D)(3|B))|((4|C)(3|B)(2|A)(5|D))|((4|C)(3|B)(5|D)(2|A))|((4|C)(5|D
)(2|A)(3|B))|((4|C)(5|D)(3|B)(2|A))|((5|D)(2|A)(3|B)(4|C))|((5|D)(
2|A)(4|C)(3|B))|((5|D)(3|B)(2|A)(4|C))|((5|D)(3|B)(4|C)(2|A))|((5|
D)(4|C)(2|A)(3|B))|((5|D)(4|C)(3|B)(2|A))|((2|A)(3|B)(9|H))|((2|A)
(3|B)(9|H))|((2|A)(9|H)(3|B))|((2|A)(9|H)(3|B))|((2|A)(3|B)(9|H))|
((2|A)(9|H)(3|B))|((3|B)(2|A)(9|H))|((3|B)(2|A)(9|H))|((3|B)(9|H)(
2|A))|((3|B)(9|H)(2|A))|((3|B)(2|A)(9|H))|((3|B)(9|H)(2|A))|((9|H)
(2|A)(3|B))|((9|H)(2|A)(3|B))|((9|H)(3|B)(2|A))|((9|H)(3|B)(2|A))|
((9|H)(2|A)(3|B))|((9|H)(3|B)(2|A))|((2|A)(3|B)(9|H))|((2|A)(9|H)(
3|B))|((3|B)(2|A)(9|H))|((3|B)(9|H)(2|A))|((9|H)(2|A)(3|B))|((9|H)
(3|B)(2|A))|((2|A)(4|C)(8|G))|((2|A)(4|C)(8|G))|((2|A)(8|G)(4|C))|
((2|A)(8|G)(4|C))|((2|A)(4|C)(8|G))|((2|A)(8|G)(4|C))|((4|C)(2|A)(
8|G))|((4|C)(2|A)(8|G))|((4|C)(8|G)(2|A))|((4|C)(8|G)(2|A))|((4|C)
(2|A)(8|G))|((4|C)(8|G)(2|A))|((8|G)(2|A)(4|C))|((8|G)(2|A)(4|C))|
((8|G)(4|C)(2|A))|((8|G)(4|C)(2|A))|((8|G)(2|A)(4|C))|((8|G)(4|C)(
2|A))|((2|A)(4|C)(8|G))|((2|A)(8|G)(4|C))|((4|C)(2|A)(8|G))|((4|C)
(8|G)(2|A))|((8|G)(2|A)(4|C))|((8|G)(4|C)(2|A))|((2|A)(5|D)(6|E)(9
|H))|((2|A)(5|D)(9|H)(6|E))|((2|A)(6|E)(5|D)(9|H))|((2|A)(6|E)(9|H
)(5|D))|((2|A)(9|H)(5|D)(6|E))|((2|A)(9|H)(6|E)(5|D))|((5|D)(2|A)(
6|E)(9|H))|((5|D)(2|A)(9|H)(6|E))|((5|D)(6|E)(2|A)(9|H))|((5|D)(6|
E)(9|H)(2|A))|((5|D)(9|H)(2|A)(6|E))|((5|D)(9|H)(6|E)(2|A))|((6|E)
(2|A)(5|D)(9|H))|((6|E)(2|A)(9|H)(5|D))|((6|E)(5|D)(2|A)(9|H))|((6
|E)(5|D)(9|H)(2|A))|((6|E)(9|H)(2|A)(5|D))|((6|E)(9|H)(5|D)(2|A))|
((9|H)(2|A)(5|D)(6|E))|((9|H)(2|A)(6|E)(5|D))|((9|H)(5|D)(2|A)(6|E
))|((9|H)(5|D)(6|E)(2|A))|((9|H)(6|E)(2|A)(5|D))|((9|H)(6|E)(5|D)(
2|A))|((2|A)(5|D)(7|F)(8|G))|((2|A)(5|D)(8|G)(7|F))|((2|A)(7|F)(5|
D)(8|G))|((2|A)(7|F)(8|G)(5|D))|((2|A)(8|G)(5|D)(7|F))|((2|A)(8|G)
(7|F)(5|D))|((5|D)(2|A)(7|F)(8|G))|((5|D)(2|A)(8|G)(7|F))|((5|D)(7
|F)(2|A)(8|G))|((5|D)(7|F)(8|G)(2|A))|((5|D)(8|G)(2|A)(7|F))|((5|D
)(8|G)(7|F)(2|A))|((7|F)(2|A)(5|D)(8|G))|((7|F)(2|A)(8|G)(5|D))|((
7|F)(5|D)(2|A)(8|G))|((7|F)(5|D)(8|G)(2|A))|((7|F)(8|G)(2|A)(5|D))
|((7|F)(8|G)(5|D)(2|A))|((8|G)(2|A)(5|D)(7|F))|((8|G)(2|A)(7|F)(5|
D))|((8|G)(5|D)(2|A)(7|F))|((8|G)(5|D)(7|F)(2|A))|((8|G)(7|F)(2|A)
(5|D))|((8|G)(7|F)(5|D)(2|A))|((2|A)(5|D))|((2|A)(5|D))|((2|A)(5|D
))|((5|D)(2|A))|((5|D)(2|A))|((5|D)(2|A))|((2|A)(5|D))|((2|A)(5|D)
)|((5|D)(2|A))|((5|D)(2|A))|((2|A)(5|D))|((5|D)(2|A))|((3|B)(3|B)(
4|C)(4|C))|((3|B)(4|C)(3|B)(4|C))|((3|B)(4|C)(4|C)(3|B))|((4|C)(3|
B)(3|B)(4|C))|((4|C)(3|B)(4|C)(3|B))|((4|C)(4|C)(3|B)(3|B))|((3|B)
(3|B)(7|F)(9|H))|((3|B)(3|B)(9|H)(7|F))|((3|B)(7|F)(3|B)(9|H))|((3
|B)(7|F)(9|H)(3|B))|((3|B)(9|H)(3|B)(7|F))|((3|B)(9|H)(7|F)(3|B))|
((7|F)(3|B)(3|B)(9|H))|((7|F)(3|B)(9|H)(3|B))|((7|F)(9|H)(3|B)(3|B
))|((9|H)(3|B)(3|B)(7|F))|((9|H)(3|B)(7|F)(3|B))|((9|H)(7|F)(3|B)(
3|B))|((3|B)(4|C)(6|E)(9|H))|((3|B)(4|C)(9|H)(6|E))|((3|B)(6|E)(4|
C)(9|H))|((3|B)(6|E)(9|H)(4|C))|((3|B)(9|H)(4|C)(6|E))|((3|B)(9|H)
(6|E)(4|C))|((4|C)(3|B)(6|E)(9|H))|((4|C)(3|B)(9|H)(6|E))|((4|C)(6
|E)(3|B)(9|H))|((4|C)(6|E)(9|H)(3|B))|((4|C)(9|H)(3|B)(6|E))|((4|C
)(9|H)(6|E)(3|B))|((6|E)(3|B)(4|C)(9|H))|((6|E)(3|B)(9|H)(4|C))|((
6|E)(4|C)(3|B)(9|H))|((6|E)(4|C)(9|H)(3|B))|((6|E)(9|H)(3|B)(4|C))
|((6|E)(9|H)(4|C)(3|B))|((9|H)(3|B)(4|C)(6|E))|((9|H)(3|B)(6|E)(4|
C))|((9|H)(4|C)(3|B)(6|E))|((9|H)(4|C)(6|E)(3|B))|((9|H)(6|E)(3|B)
(4|C))|((9|H)(6|E)(4|C)(3|B))|((3|B)(4|C)(7|F)(8|G))|((3|B)(4|C)(8
|G) (7|F))|((3|B) (7|F) (4|C) (8|G))|((3|B) (7|F) (8|G) (4|C))|((3|B) (8|G
) (4|C) (7|F))|((3|B) (8|G) (7|F) (4|C))|((4|C) (3|B) (7|F) (8|G))|((4|C) (
3|B) (8|G) (7|F))|((4|C) (7|F) (3|B) (8|G))|((4|C) (7|F) (8|G) (3|B))|((4|
C) (8|G) (3|B) (7|F))|((4|C) (8|G) (7|F) (3|B))|((7|F) (3|B) (4|C) (8|G))|(
(7|F) (3|B) (8|G) (4|C))|((7|F) (4|C) (3|B) (8|G))|((7|F) (4|C) (8|G) (3|B)
)|((7|F) (8|G) (3|B) (4|C))|((7|F) (8|G) (4|C) (3|B))|((8|G) (3|B) (4|C) (7
|F))|((8|G) (3|B) (7|F) (4|C))|((8|G) (4|C) (3|B) (7|F))|((8|G) (4|C) (7|F
) (3|B))|((8|G) (7|F) (3|B) (4|C))|((8|G) (7|F) (4|C) (3|B))|((3|B) (4|C))
|((3|B) (4|C))|((3|B) (4|C))|((4|C) (3|B))|((4|C) (3|B))|((4|C) (3|B))|
((3|B) (4|C))|((3|B) (4|C))|((4|C) (3|B))|((4|C) (3|B))|((3|B) (4|C))|(
(4|C) (3|B))|((3|B) (5|D) (7|F))|((3|B) (5|D) (7|F))|((3|B) (7|F) (5|D))|
((3|B) (7|F) (5|D))|((3|B) (5|D) (7|F))|((3|B) (7|F) (5|D))|((5|D) (3|B) (
7|F))|((5|D) (3|B) (7|F))|((5|D) (7|F) (3|B))|((5|D) (7|F) (3|B))|((5|D)
(3|B) (7|F))|((5|D) (7|F) (3|B))|((7|F) (3|B) (5|D))|((7|F) (3|B) (5|D))|
((7|F) (5|D) (3|B))|((7|F) (5|D) (3|B))|((7|F) (3|B) (5|D))|((7|F) (5|D) (
3|B))|((3|B) (5|D) (7|F))|((3|B) (7|F) (5|D))|((5|D) (3|B) (7|F))|((5|D)
(7|F) (3|B))|((7|F) (3|B) (5|D))|((7|F) (5|D) (3|B))|((4|C) (4|C) (6|E) (8
|G))|((4|C) (4|C) (8|G) (6|E))|((4|C) (6|E) (4|C) (8|G))|((4|C) (6|E) (8|G
) (4|C))|((4|C) (8|G) (4|C) (6|E))|((4|C) (8|G) (6|E) (4|C))|((6|E) (4|C) (
4|C) (8|G))|((6|E) (4|C) (8|G) (4|C))|((6|E) (8|G) (4|C) (4|C))|((8|G) (4|
C) (4|C) (6|E))|((8|G) (4|C) (6|E) (4|C))|((8|G) (6|E) (4|C) (4|C))|((4|C)
(5|D) (6|E))|((4|C) (5|D) (6|E))|((4|C) (6|E) (5|D))|((4|C) (6|E) (5|D))|
((4|C) (5|D) (6|E))|((4|C) (6|E) (5|D))|((5|D) (4|C) (6|E))|((5|D) (4|C) (
6|E))|((5|D) (6|E) (4|C))|((5|D) (6|E) (4|C))|((5|D) (4|C) (6|E))|((5|D)
(6|E) (4|C))|((6|E) (4|C) (5|D))|((6|E) (4|C) (5|D))|((6|E) (5|D) (4|C))|
((6|E) (5|D) (4|C))|((6|E) (4|C) (5|D))|((6|E) (5|D) (4|C))|((4|C) (5|D) (
6|E))|((4|C) (6|E) (5|D))|((5|D) (4|C) (6|E))|((5|D) (6|E) (4|C))|((6|E)
(4|C) (5|D))|((6|E) (5|D) (4|C))|((5|D) (5|D) (6|E) (7|F))|((5|D) (5|D) (7
|F) (6|E))|((5|D) (6|E) (5|D) (7|F))|((5|D) (6|E) (7|F) (5|D))|((5|D) (7|F
) (5|D) (6|E))|((5|D) (7|F) (6|E) (5|D))|((6|E) (5|D) (5|D) (7|F))|((6|E) (
5|D) (7|F) (5|D))|((6|E) (7|F) (5|D) (5|D))|((7|F) (5|D) (5|D) (6|E))|((7|
F) (5|D) (6|E) (5|D))|((7|F) (6|E) (5|D) (5|D))|((6|E) (6|E) (9|H) (9|H))|(
(6|E) (9|H) (6|E) (9|H))|((6|E) (9|H) (9|H) (6|E))|((9|H) (6|E) (6|E) (9|H)
)|((9|H) (6|E) (9|H) (6|E))|((9|H) (9|H) (6|E) (6|E))|((6|E) (7|F) (8|G) (9
|H))|((6|E) (7|F) (9|H) (8|G))|((6|E) (8|G) (7|F) (9|H))|((6|E) (8|G) (9|H
) (7|F))|((6|E) (9|H) (7|F) (8|G))|((6|E) (9|H) (8|G) (7|F))|((7|F) (6|E) (
8|G) (9|H))|((7|F) (6|E) (9|H) (8|G))|((7|F) (8|G) (6|E) (9|H))|((7|F) (8|
G) (9|H) (6|E))|((7|F) (9|H) (6|E) (8|G))|((7|F) (9|H) (8|G) (6|E))|((8|G)
(6|E) (7|F) (9|H))|((8|G) (6|E) (9|H) (7|F))|((8|G) (7|F) (6|E) (9|H))|((8
|G) (7|F) (9|H) (6|E))|((8|G) (9|H) (6|E) (7|F))|((8|G) (9|H) (7|F) (6|E))|
((9|H) (6|E) (7|F) (8|G))|((9|H) (6|E) (8|G) (7|F))|((9|H) (7|F) (6|E) (8|G
))|((9|H) (7|F) (8|G) (6|E))|((9|H) (8|G) (6|E) (7|F))|((9|H) (8|G) (7|F) (
6|E))|((6|E) (9|H))|((6|E) (9|H))|((6|E) (9|H))|((9|H) (6|E))|((9|H) (6
|E))|((9|H) (6|E))|((6|E) (9|H))|((6|E) (9|H))|((9|H) (6|E))|((9|H) (6|
E))|((6|E) (9|H))|((9|H) (6|E))|((7|F) (7|F) (8|G) (8|G))|((7|F) (8|G) (7
|F) (8|G))|((7|F) (8|G) (8|G) (7|F))|((8|G) (7|F) (7|F) (8|G))|((8|G) (7|F
) (8|G) (7|F))|((8|G) (8|G) (7|F) (7|F))|((7|F) (8|G))|((7|F) (8|G))|((7|
F) (8|G))|((8|G) (7|F))|((8|G) (7|F))|((8|G) (7|F))|((7|F) (8|G))|((7|F
) (8|G))|((8|G) (7|F))|((8|G) (7|F))|((7|F) (8|G))|((8|G) (7|F))
```