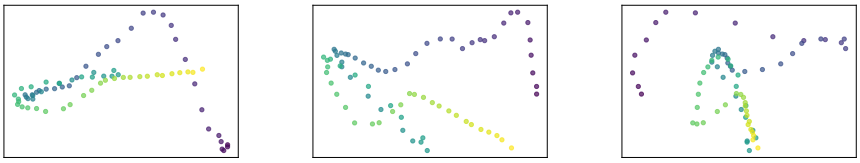
A APPENDIX - SECTION 1

Table 1: Hyperparameters for the experiments reported in the hard exploration table.

Hyperparameter	Symbol	Value (MontezumaRevenge, Gravitar, Frostbite)
Learning rate	lr	$1e^{-4}$
Batch size	bs	64
Time window	L	$(20, 25, 25)$
Distance threshold	T_d	$(0.6, 0.75, 0.75)$
Visit threshold	T_v	0.65
Proxy cell interval	$[T_{\text{p-low}}, T_{\text{p-high}}]$	$[0.45, 0.75]$
Exploration steps	t	40
action repetition mean	μ	4

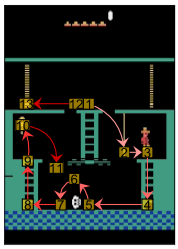


(a) dim 1-2 (b) dim 2-3 (c) dim 1-3

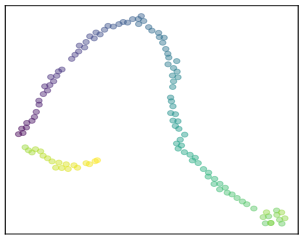
Figure 1: Principal Component Analysis for the trajectory visualization.

Algorithm 1 Go-Explore with a learned state representation

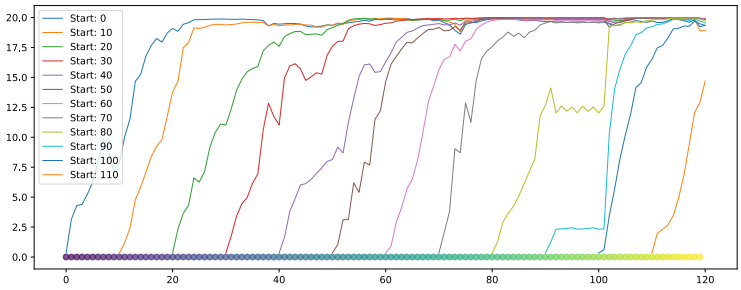
Initialize archive, dataset, agent, network
for iteration $= 1, 2, ...$ **do**
  Let the agent act t timesteps in the environment starting from the selected and proxy cells
  Collect data and optimize L^{time} w.r.t. θ
  Recompute all archive cell representations: $z_C = \Phi_\theta(\bar{o}_C)$
  for each trajectory $= 1, 2, ..., N$ **do**
    Transfer all observations $\bar{o}_1, ..., \bar{o}_t$ into the latent representation $z_1, ..., z_t$
    Compute all necessary time distances $\Psi_\theta(z_c, z_{1,...,t})$
    Apply archive insertion criterion to a candidate w.r.t. threshold T_d
    Increase cell visits w.r.t. threshold T_v
    Collect some proxy cells w.r.t. threshold $[T_{\text{p-low}}, T_{\text{p-high}}]$
    if candidate is accepted **then**
      Add candidate to archive
    end if
  end for
end for



(a) Longer human demonstration that reaches 400 score.



(b) t-SNE visualization.



(c) Time prediction capability on the shown trajectory (a).

Figure 2: Sophisticated model and its capabilities on a longer trajectory. The shown model is trained on data that surpasses the shown trajectory (a). Again, we can see the good encoding property and an improved time prediction skill. The predicted times around the timesteps 90-100 or in the image at the Box 11 are accurate. At that point the agent dies and the environment generates repeating frames (two-image sprites) for around 10 frames. This prediction behavior happens, because we remove temporal ambiguity between state-pairs and try to calculate the shortest distance for it. The event can also be seen in the t-SNE visualization, when the light-green points start to form a cluster.

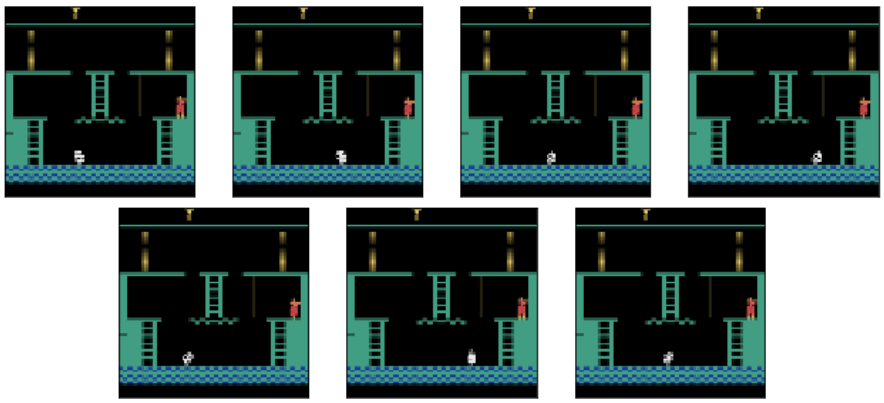


Figure 3: Creation of archive entries. We manually looked through the cell observations of an archive and searched for similarity. This figure shows *all* cell observations where the agent stands at the same position and already collected the key. We can see that the time prediction network does not allow duplicates in the archive and keeps a reasonable distance between the observations where the white skull is changing positions. Moreover the archive holds no observation where the agent just slightly moved in these situations.