A Implementation Details

We would like to preface the appendix with a discussion on the relative differences in the assumptions that Q*BERT and MC!Q*BERT make regarding the underlying environment. Although both are framed as POMDPs, MC!Q*BERT makes stronger assumptions regarding the determinism of the game as compared to Q*BERT. MC!Q*BERT (and GO!Q*BERT) rely on the fact that the set of transition probabilities in a text-game are mostly deterministic. Using this, they are able to assume that frozen policies can be executed deterministically, i.e. with no significant deviations from the original trajectory. It is possible to robustify such policies by extending our method of structured exploration to perhaps perform imitation learning on the found highest score trajectories as seen in Phase 2 of the original GoExplore algorithm Ecoffet et al. (2019). Stochasticity is not among set of challenges tackled in this work, however—we focus on learning how to better explore combinatorially-sized spaces with underlying long-term dependencies. For future works in this space, we believe that agents should be compared based on the set of assumptions made: agents like KG-A2C and Q*BERT when operating under standard reinforcement learning assumptions, and MC!Q*BERT and GO!Q*BERT when under the stronger assumption of having a deterministic simulator.

A.1 Q*BERT

This section outlines how Q*BERT is trained, including details of the Jericho-QA dataset, the overall architecture, A2C training and hyperparameter details.

A.1.1 Jericho-QA Dataset

Jericho-QA contains 221453 Question-Answer pairs in the training set and 56667 pairs in the held out test set. The test set consists of all the games that we test on in this paper. It is collected by randomly exploring games using a set of admissible actions in addition to using the walkthroughs for each game as found in the Jericho framework (Hausknecht et al., 2020). The set of attributes for a game is taken directly from the game engine and is defined by the game developer.

Questions for QA were chosen on the basis of what past works in the area determined to be useful state information for the agent. For example, KG-A2C and GATA (Adhikari et al., 2020) explicitly differentiate between inventory/location descriptions/surrounding objects. The answers are annotated using information in the underlying world object tree that every text game is built on, this information can be accessed through the engine but, importantly, is used for annotation in the Jericho-QA dataset only. Jericho-QA data is formatted in the style of SQuAD 2.0 (Rajpurkar et al., 2018) and given samples of which questions are not applicable to certain states, i.e. negative samples. All games in the Jericho suite except the 9 games tested on in this paper were used to collect training data.

A single sample looks like this:

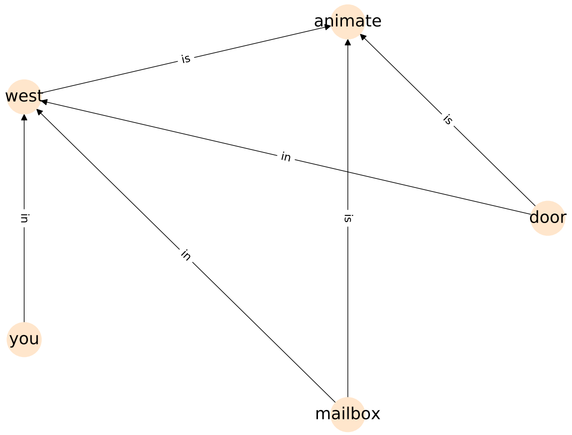
```
Context:
[loc] Chief’s Office You are standing in the chief’s office. He is telling you, "The mayor was murdered
      yeaterday night at 12:03 am. I want you to solve it before we get any bad publicity or the FBI has to
      come in." "Yessir!" you reply. He hands you a sheet of paper. once you have read it, go north or west.
      You can see a piece of white paper here.
[inv] You are carrying nothing.
[obs] [your score has just gone up by ten points.]
[atr] talkable, seen, lieable, enterable, nodwarf, indoors, visited, handed, lockable, surface, thing,
      water_room, unlock, lost, afflicted, is_treasure, converse, mentioned, male, npcworn, no_article,
      relevant, scored, queryable, town, pluggable, happy, is_followable, legible, multitude, burning, room,
      clothing, underneath, ward_area , little, intact, animate, bled_in, supporter, readable, openable, near,
       nonlocal, door, plugged, sittable, toolbit, vehicle, light, lens_searchable, open, familiar, is_scroll,
       aimable, takeable, static, unique, concealed, vowelstart, alcoholic, bodypart, general, is_spell, full,
       dry_land, pushable, known, proper, inside, clean, ambiguously_plural, container, edible, treasure,
      can_plug, weapon, is_arrow, insubstantial, pluralname, transparent, is_coin, air_room, scenery, on,
      is_spell_book, burnt, burnable, auto_searched, locked, switchable, absent, rockable, beenunlocked,
      progressing, severed, worn, windy, stone, random, neuter, legible, female, asleep, wiped
Question: Where am I located? Answer: chief’s office
Question: What is here? Answer: paper, west
Question: What do I have? Answer: nothing
Question: What attributes does paper have? Answer: legible, animate
Question: What attributes does west have? Answer: room, animate
```

A.1.2 Knowledge Graph Update Rules

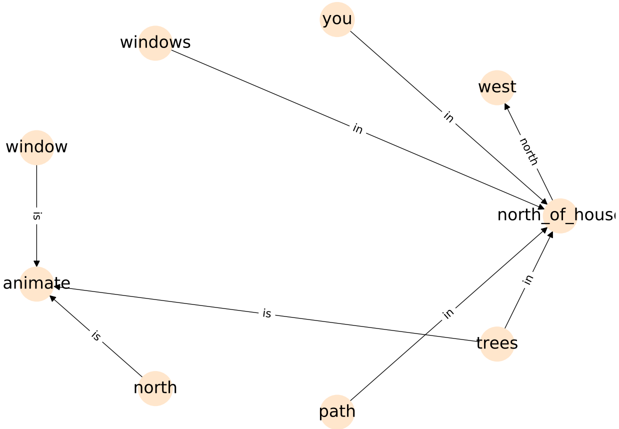
Every step, given the current state and possible attributes as context—the QA network predicts the current room location, the set of all inventory objects, the set of all surrounding objects, and all attributes for each object.

- Linking the current room type (e.g. “Kitchen”, “Cellar”) to the items found in the room with the relation “has”, e.g. $\langle kitchen, has, lamp\rangle$
- All attribute information for each object is linked to the object with the relation “is”. e.g. $\langle egg, is, treasure\rangle$
- Linking all inventory objects with relation “have” to the “you” node, e.g. $\langle you, have, sword\rangle$
- Linking rooms with directions based on the action taken to move between the rooms, e.g. $\langle Behind\ House, east\ of, Forest\rangle$ after the action “go east” is taken to go from behind the house to the forest

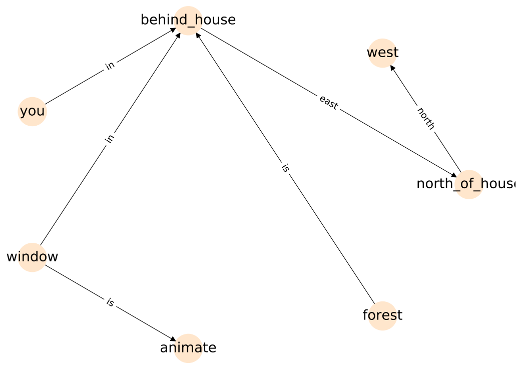
Below is an excerpt from *Zork1* showing the exact observations given to the Q*BERT,the knowledge graph, and the corresponding action taken by the agent after the graph extraction and update process has occurred as described above for a trajectory consisting of 5 timesteps. These timesteps begin at the start of the game in *West of House* and continue till the agent has entered the *Kitchen* as seen in Fig. 2 and Fig. 7. The set of $\langle s, r, o\rangle$ triples that make up the graph are in the text and the figure shows a partial visualization of the graph at that particular step in the trajectory.



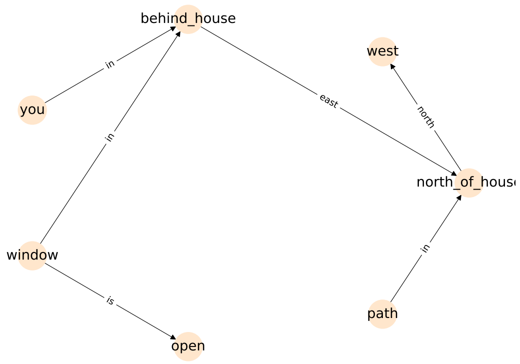
```
[loc] West of House You are standing in an open field west of a white house, with a boarded front door. There
      is a small mailbox here. [inv] You are empty handed.
[obs] Copyright c 1981, 1982, 1983 Infocom, Inc. All rights reserved. ZORK is a registered trademark of
      Infocom, Inc. Revision 88 / Serial number 840726 West of House You are standing in an open field west of
       a white house, with a boarded front door. There is a small mailbox here.
[atr] talkable, seen, lieable, enterable, nodwarf, indoors, visited, handed, lockable, surface, thing,
      water_room, unlock, lost, afflicted, is_treasure, converse, mentioned, male, npcworn, no_article,
      relevant, scored, queryable, town, pluggable, happy, is_followable, legible, multitude, burning, room,
      clothing, underneath, ward_area, little, intact, animate, bled_in, supporter, readable, openable, near,
      nonlocal, door, plugged, sittable, toolbit, vehicle, light, lens_searchable, open, familiar, is_scroll,
      aimable, takeable, static, unique, concealed, vowelstart, alcoholic, bodypart, general, is_spell, full,
      dry_land, pushable, known, proper, inside, clean, ambiguously_plural, container, edible, treasure,
      can_plug, weapon, is_arrow, insubstantial, pluralname, transparent, is_coin, air_room, scenery, on,
      is_spell_book, burnt, burnable, auto_searched, locked, switchable, absent, rockable, beenunlocked,
      progressing, severed, worn, windy, stone, random, neuter, legible, female, asleep, wiped
[graph] [(you, in, west), (door, is, animate), (door, in, west), (west, is, animate), (mailbox, in, west), (
      mailbox, is, animate)]
[next act] go north
```



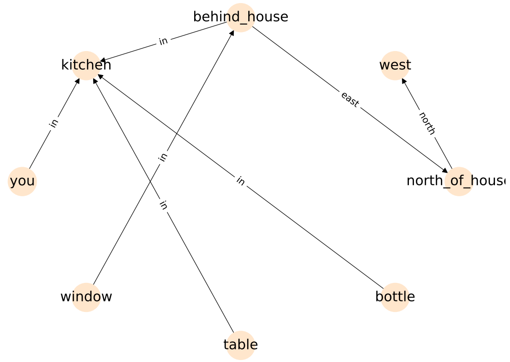
```
[loc] North of House You are facing the north side of a white house. There is no door here, and all the
      windows are boarded up. To the north a narrow path winds through the trees.
[inv] You are empty handed.
[obs] North of House You are facing the north side of a white house. There is no door here, and all the
      windows are boarded up. To the north a narrow path winds through the trees.
[atr] talkable, seen, lieable, enterable, nodwarf, indoors, visited, handed, lockable, surface, thing,
      water_room, unlock, lost, afflicted, is_treasure, converse, mentioned, male, npcworn, no_article,
      relevant, scored, queryable, town, pluggable, happy, is_followable, legible, multitude, burning, room,
      clothing, underneath, ward_area, little, intact, animate, bled_in, supporter, readable, openable, near,
      nonlocal, door, plugged, sittable, toolbit, vehicle, light, lens_searchable, open, familiar, is_scroll,
      aimable, takeable, static, unique, concealed, vowelstart, alcoholic, bodypart, general, is_spell, full,
      dry_land, pushable, known, proper, inside, clean, ambiguously_plural, container, edible, treasure,
      can_plug, weapon, is_arrow, insubstantial, pluralname, transparent, is_coin, air_room, scenery, on,
      is_spell_book, burnt, burnable, auto_searched, locked, switchable, absent, rockable, beenunlocked,
      progressing, severed, worn, windy, stone, random, neuter, legible, female, asleep, wiped
[graph] [(north_of_house, north, west), (you, in, north_of_house), (door, is, animate), (door, in, west), (
      west, is, animate), (west, in, west), (mailbox, in, west), (mailbox, is, animate), (windows, in,
      north_of_house), (windows, is, animate), (north, is, animate), (north, in, north_of_house), (path, is,
      animate), (path, in, north_of_house), (trees, in, north_of_house), (trees, is, animate)]
[next act] go east
```



```
[loc] Behind House You are behind the white house. A path leads into the forest to the east. In one corner of
     the house there is a small window which is slightly ajar.
[inv] You are empty handed.
[obs] Behind House You are behind the white house. A path leads into the forest to the east. In one corner of
     the house there is a small window which is slightly ajar.
[atr] talkable, seen, lieable, enterable, nodwarf, indoors, visited, handed, lockable, surface, thing,
     water_room, unlock, lost, afflicted, is_treasure, converse, mentioned, male, npcworn, no_article,
     relevant, scored, queryable, town, pluggable, happy, is_followable, legible, multitude, burning, room,
     clothing, underneath, ward_area, little, intact, animate, bled_in, supporter, readable, openable, near,
     nonlocal, door, plugged, sittable, toolbit, vehicle, light, lens_searchable, open, familiar, is_scroll,
     aimable, takeable, static, unique, concealed, vowelstart, alcoholic, bodypart, general, is_spell, full,
     dry_land, pushable, known, proper, inside, clean, ambiguously_plural, container, edible, treasure,
     can_plug, weapon, is_arrow, insubstantial, pluralname, transparent, is_coin, air_room, scenery, on,
     is_spell_book, burnt, burnable, auto_searched, locked, switchable, absent, rockable, beenunlocked,
     progressing, severed, worn, windy, stone, random, neuter, legible, female, asleep, wiped
[graph] [(north_of_house, north, west), (behind_house, east, north_of_house), (you, in, behind_house), (door,
     is, animate), (door, in, west), (west, is, animate), (west, in, west), (you, in, behind_house), (mailbox,
      in, west), (mailbox, is, animate), (windows, in, north_of_house), (windows, is, animate), (north, is,
     animate), (north, in, north_of_house), (path, is, animate), (path, in, north_of_house), (trees, in,
     north_of_house), (trees, is, animate), (window, in, behind_house), (window, is, animate), (forest, in,
     behind_house), (forest, is, animate), (east, in, behind_house), (east, is, animate)]
[next act] open window
```



```
[loc] Behind House You are behind the white house. A path leads into the forest to the east. In one corner of
     the house there is a small window which is open. [inv] You are empty handed.
[obs] With great effort, you open the window far enough to allow entry.
[atr] talkable, seen, lieable, enterable, nodwarf, indoors, visited, handed, lockable, surface, thing,
     water_room, unlock, lost, afflicted, is_treasure, converse, mentioned, male, npcworn, no_article,
     relevant, scored, queryable, town, pluggable, happy, is_followable, legible, multitude, burning, room,
     clothing, underneath, ward_area, little, intact, animate, bled_in, supporter, readable, openable, near,
     nonlocal, door, plugged, sittable, toolbit, vehicle, light, lens_searchable, open, familiar, is_scroll,
     aimable, takeable, static, unique, concealed, vowelstart, alcoholic, bodypart, general, is_spell, full,
     dry_land, pushable, known, proper, inside, clean, ambiguously_plural, container, edible, treasure,
     can_plug, weapon, is_arrow, insubstantial, pluralname, transparent, is_coin, air_room, scenery, on,
     is_spell_book, burnt, burnable, auto_searched, locked, switchable, absent, rockable, beenunlocked,
     progressing, severed, worn, windy, stone, random, neuter, legible, female, asleep, wiped
[graph] [(north_of_house, north, west), (behind_house, east, north_of_house), (you, in, behind_house), (door,
     is, animate), (door, in, west), (west, is, animate), (west, in, west), (mailbox, in, west), (mailbox, is,
      animate), (windows, in, north_of_house), (windows, is, animate), (windows, is, open), (north, is,
     animate), (north, in, north_of_house), (path, is, animate), (path, in, north_of_house), (trees, in,
     north_of_house), (trees, is, animate), (window, in, behind_house), (window, is, animate), (forest, in,
     behind_house), (forest, is, animate), (east, in, behind_house), (east, is, animate)]
```



```
[loc] Kitchen You are in the kitchen of the white house. A table seems to have been used recently for the
      preparation of food. A passage leads to the west and a dark staircase can be seen leading upward. A dark
       chimney leads down and to the east is a small window which is open. On the table is an elongated brown
      sack, smelling of hot peppers. A bottle is sitting on the table. The glass bottle contains: A quantity
      of water
[inv] You are empty handed.
[obs] Kitchen You are in the kitchen of the white house. A table seems to have been used recently for the
      preparation of food. A passage leads to the west and a dark staircase can be seen leading upward. A dark
       chimney leads down and to the east is a small window which is open. On the table is an elongated brown
      sack, smelling of hot peppers. A bottle is sitting on the table. The glass bottle contains: A quantity
      of water
[atr] talkable, seen, lieable, enterable, nodwarf, indoors, visited, handed, lockable, surface, thing,
      water_room, unlock, lost, afflicted, is_treasure, converse, mentioned, male, npcworn, no_article,
      relevant, scored, queryable, town, pluggable, happy, is_followable, legible, multitude, burning, room,
      clothing, underneath, ward_area, little, intact, animate, bled_in, supporter, readable, openable, near,
      nonlocal, door, plugged, sittable, toolbit, vehicle, light, lens_searchable, open, familiar, is_scroll,
      aimable, takeable, static, unique, concealed, vowelstart, alcoholic, bodypart, general, is_spell, full,
      dry_land, pushable, known, proper, inside, clean, ambiguously_plural, container, edible, treasure,
      can_plug, weapon, is_arrow, insubstantial, pluralname, transparent, is_coin, air_room, scenery, on,
      is_spell_book, burnt, burnable, auto_searched, locked, switchable, absent, rockable, beenunlocked,
      progressing, severed, worn, windy, stone, random, neuter, legible, female, asleep, wiped
[graph] [(north_of_house, north, west), (behind_house, east, north_of_house), (behind_house, in, kitchen), (
      you, in, kitchen), (door, is, animate), (door, in, west), (west, is, animate), (west, in, west), (west,
      in, kitchen), (mailbox, in, west), (mailbox, is, animate), (windows, in, north_of_house), (windows, is,
      animate), (north, is, animate), (north, in, north_of_house), (path, is, animate), (path, in,
      north_of_house), (trees, in, north_of_house), (trees, is, animate), (window, in, behind_house), (window,
       is, animate), (forest, in, behind_house), (forest, is, animate), (east, in, behind_house), (east, is,
      animate), (table, in, kitchen), (table, is, animate)]]
[next act] go in
```

A.1.3 Architecture

Further details of what is found in Figure 3. The sequential action decoder consists two GRUs that are linked together as seen in Ammanabrolu & Hausknecht (2020). The first GRU decodes an action template and the second decodes objects that can be filled into the template. These objects are constrained by a *graph mask*, i.e. the decoder is only allowed to select entities that are already present in the knowledge graph.

The question answering network based on ALBERT (Lan et al., 2020) has the following hyperparameters, taken from the original paper and known to work well on the SQuAD 2.0 (Rajpurkar et al., 2018) dataset. No further hyperparameter tuning was conducted.

Parameters	**Value**
batch size	8
learning rate	3×10^{-5}
max seq len	512
doc stride	128
warmup steps	814
max steps	8144
gradient accumulation steps	24

A.1.4 A2C Training

The rest of the A2C training is unchanged from Ammanabrolu & Hausknecht (2020).

Action Space Solving *Zork1*, the canonical text-adventure game, requires the generation of actions consisting of up to five-words from a relatively modest vocabulary of 697 words recognized by the game's parser. This results in $\mathcal{O}(697^5) = 1.64 \times 10^{14}$ possible actions at every step. Hausknecht et al. (2020) propose a template-based action space in which the agent first selects a template, consisting of an action verb and preposition, and then filling that in with relevant entities (e.g. $[get]$ $\_\_[from]\_\_$). *Zork1* has 237 templates, each with up to two blanks, yielding a template-action space of size $\mathcal{O}(237 \times 697^2) = 1.15 \times 10^8$. This space is still far larger than most used by previous approaches applying reinforcement learning to text-based games. We use this template action space for all games.

A2C training starts with calculating the advantage of taking an action in a state $A(s_t, a_t)$, defined as the value of taking an action $Q(s_t, a_t)$ compared to the average value of taking all possible *admissible actions* in that state $V(s_t)$:

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t) \tag{5}$$

$$Q(s_t, a_t) = \mathbb{E}[r_t + \gamma V(s_{t+1})] \tag{6}$$

The value is predicted by the critic as shown in Fig. 3 and r_t is the reward received at step t.

The action decoder or actor is then updated according to the gradient:

$$-\nabla_\theta(log\pi_{\mathbb{T}}(\tau|s_t;\theta_t) + \sum_{i=1}^{n} log\pi_{\mathbb{O}_i}(o_i|s_t,\tau,...,o_{i-1};\theta_t))A(s_t,a_t) \tag{7}$$

updating the template policy $\pi_{\mathbb{T}}$ and object policies $\pi_{\mathbb{O}_i}$ based on the fact that each step in the action decoding process is conditioned on all the previously decoded portions. The critic is updated with respect to the gradient:

$$\frac{1}{2}\nabla_\theta(Q(s_t,a_t;\theta_t) - V(s_t;\theta_t))^2 \tag{8}$$

bringing the critic's prediction of the value of being in a state closer to its true underlying value. An entropy loss is also added:

$$\mathcal{L}_{\mathbb{E}}(s_t,a_t;\theta_t) = \sum_{a \in V(s_t)} P(a|s_t)logP(a|s_t) \tag{9}$$

Hyperparameters are taken from KG-A2C as detailed by Ammanabrolu & Hausknecht (2020) and not tuned any further.

A.2 MC!Q*BERT

The additional hyperparamters used for modular policy chaining are detailed below. *Patience batch factor* is the proportion of the batch that must have stagnated at a particular score for *patience* number of episodes of unchanging score before a bottleneck is detected. *Patience* within a range of $1000 - 6000$ in increments of 500 and *buffer size* within a range of $10 - 60$ in increments of 10 were the only additional parameters tuned for, on *Zork1*. The resulting best hyperparameter set was used on the rest of the games.

Parameters	**Value**
patience	3000
buffer size	40
batch size	16
patience batch factor	.75

A.3 GO!Q*BERT

The Go-Explore algorithm consists of two phases, the first to continuously explore until a set of promising states and corresponding trajectories are found on the basis of total score, and the second to robustify this found policy against potential stochasticity in the game. Promising states are defined as those states when explored from will likely result in higher reward trajectories. Madotto et al. (2020) look at applying Go-Explore to text-games on a set of simpler games generated using the game generation framework TextWorld (Côté et al., 2018). They use a small set of "admissible actions"—actions guaranteed to change the world state at any given step during Phase 1—to explore and find high reward trajectories.

Since the text games we are dealing with are mostly deterministic, with the exception of *Zork1* in later stages, we only focus on using Phase 1 of the Go-Explore algorithm to find an optimal policy. Go-Explore maintains an archive of cells—defined as a set of states that map to a single representation—to keep track of promising states. Ecoffet et al. (2019) simply encodes each cell by keeping track of the agent's position and Madotto et al. (2020) use the textual observations encoded by recurrent neural network as a cell representation. We improve on this implementation by training the Q*BERT network in parallel, using the snapshot of the knowledge graph in conjunction with the game state to further encode the current state and use this as a cell representation. At each step, Go-Explore chooses a cell to explore at random (weighted by score to prefer more advanced cells). Q*BERT will run for a number of steps in each cell, for all our experiments we use a cell step size of 32, starting with the knowledge graph state and the last seen state of the game from the cell. This will generate a trajectory for the agent while further training Q*BERT at each iteration, creating a new representation for the knowledge graph as well as a new game state for the cell. After expanding a cell, Go-Explore will continue to sample cells by weight to continue expanding its known states. At the same time, Q*BERT will benefit from the heuristics of selecting preferred cells and be trained on promising states more often.

B Results

B.1 Graph Evaluation Results

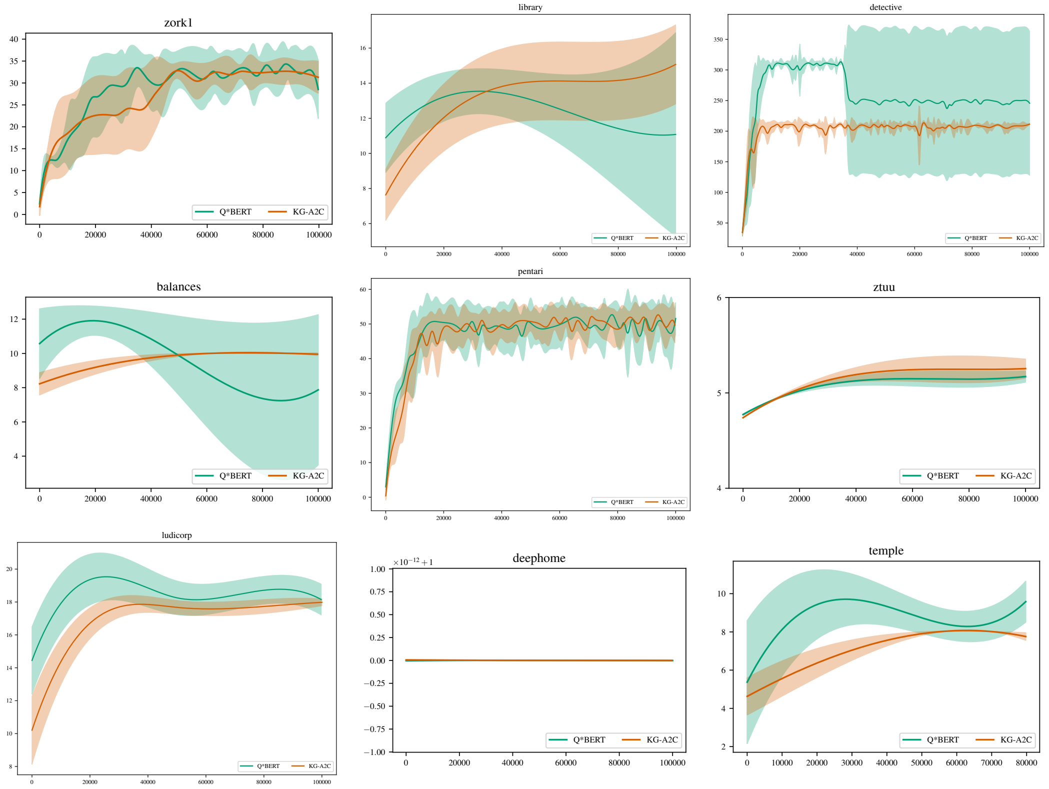


Figure 5: Episode initial reward curves for KG-A2C and Q*BERT.

B.2 Ground Truth Graph Evaluations

Here we present results for the agents when given the ground truth knowledge graphs directly from the game engine. We see marginally greater performance across the board when compared to agents using constructed knowledge graphs (seen in Table 1). This shows once again that knowledge graph accuracy is correlated to game performance, though the lower margin indicates that after a certain point—i.e. the accuracy levels of Q*BERT—gains in knowledge graph accuracy provide diminishing returns with respect to overall performance for this particular architecture.

Expt.	**Q*BERT**		**MC!Q*BERT**
Game Reward	✓		✓
Intrinsic Motive			✓
Metric	**Eps.**	**Max**	**Max**
zork1	34.5	35	42
library	4.5	18	19
detective	246.1	288	338
balances	9.8	10	10
pentari	52.7	56	58
ztuu	5	5	5
ludicorp	17.9	18	18
deephome	1	1	6
temple	8.0	8	8

Table 2: Ground truth knowledge graph experiment results.

B.3 Intrinsic Motivation and Structured Exploration Results

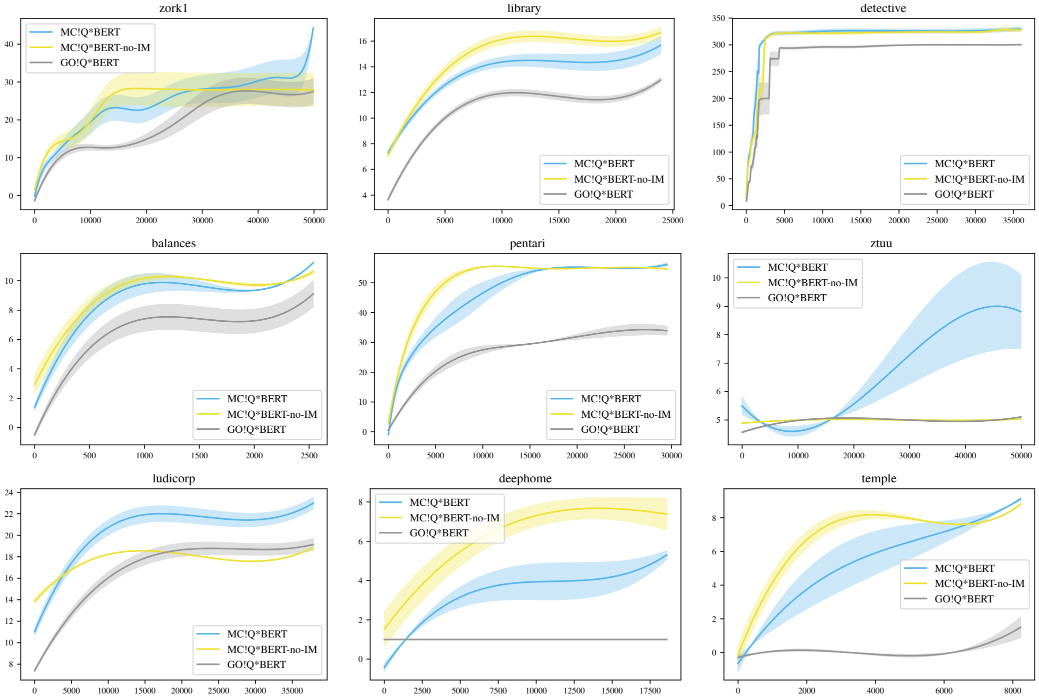


Figure 6: Best initial reward curves for the exploration strategies.

C Zork1

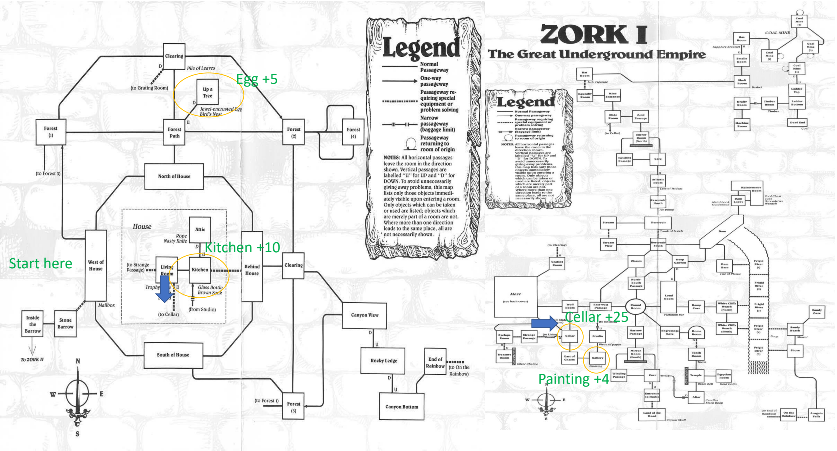


Figure 7: Map of *Zork1* annotated with rewards taken from Ammanabrolu & Hausknecht (2020) and corresponding to the states and rewards found in Figure 2.