A Appendix

A.1 Clarification on reference to [40]

As we make a reference to [40] – a work that was later retracted due to an error – we would like to clarify that we are aware of the error and our claims remain valid despite the error in cited work.

In [40], the authors mistakenly viewed BERT[9] as a Markov Random Field (MRF). While the original goal of [40] was to derive a procedure for sampling from masked language models (MLMs) by viewing them as MRFs, the work also inspired the use of MLM prediction error as a way for scoring sequences when decoding from a language model. The way we propose to use MLMs is similar to the latter, i.e. as a proxy metric for *scoring* sequence-level predictions of trajectory information. In other words, we do not formally treat the MLM in MIMEx as an MRF, and we do not attempt to obtain conditional distributions from which one generates samples. We also note that a correct energy-based view was later proposed in [13], which does not change the argument that we put forth either. While sampling from an energy-based model is expensive, we only seek to obtain a useful stochastic estimate of the energy function for the purpose of scoring.

A.2 Detailed reward design of the original PixMC tasks

We present the detailed reward terms of each task in the original PixMC. The total environment reward of each task is the sum of all reward terms, each multiplied with a tunable scale parameter.

FrankaReach: distance to goal (from parallel gripper); goal bonus when distance to goal is smaller than a threshold value; action penalty.

KukaReach: distance to goal (from humanoid hand); goal bonus when distance to goal is smaller than a threshold value; action penalty.

FrankaCabinet: distance to handle (of cabinet); handle bonus when mesh of parallel gripper intersects mesh of handle; open bonus when cabinet is open; distance to goal (from parallel gripper); open pose bonus when parallel gripper is within a certain pose distribution; goal bonus when distance to goal is smaller than a threshold value; action penalty.

KukaCabinet: distance to handle (of cabinet); handle bonus when mesh of humanoid hand intersects mesh of handle; open bonus when cabinet is open; distance to goal (from humanoid hand); open pose bonus when humanoid hand is within a certain pose distribution; goal bonus when distance to goal is smaller than a threshold value; action penalty.

FrankaPick: distance to object (from parallel gripper); lift bonus when object is lifted above the table; distance to goal (from parallel gripper); goal bonus when distance to goal is smaller than a threshold value; action penalty.

KukaPick: distance to object (from humanoid hand); lift bonus when object is lifted above the table; distance to goal (from humanoid hand); goal bonus when distance to goal is smaller than a threshold value; action penalty.

FrankaMove: distance to object (from parallel gripper); lift bonus when object is lifted above the table; distance to goal (from parallel gripper); goal bonus when distance to goal is smaller than a threshold value; action penalty.

KukaMove: distance to object (from humanoid hand); lift bonus when object is lifted above the table; distance to goal (from humanoid hand); goal bonus when distance to goal is smaller than a threshold value; action penalty.

Due to the dense reward terms, these tasks are not sufficiently challenging for benchmarking state-of-the-art exploration algorithms.

A.3 Detailed comparison between PixMC and PixMC-Sparse tasks

We visualize partial trajectory of all PixMC-Sparse tasks in Figure 6. Below, we provide more details on how agents are rewarded differently in PixMC-Sparse compared to in PixMC.

For **Reach** tasks, agent needs to move its end effector (parallel gripper in the case of Franka; humanoid hand in the case of Kuka) such that the end effector reaches a specific goal location. In PixMC, agent receives a variable-valued reward at every time step; the reward value is continuous and inversely proportional to end effector's distance to goal location. In PixMC-Sparse, agent only receives a

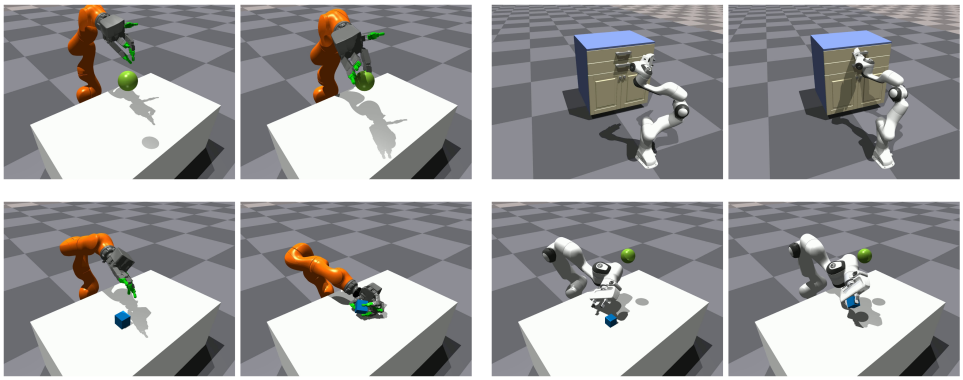

Figure 6: **Visualization of tasks.** (*Top-left*) KukaReach. (*Top-right*) FrankaCabinet. (*Bottom-left*) KukaPick. (*Bottom-right*) FrankaMove.

fixed-valued reward when its end effector intersects with a small spherical space (visualized as the green sphere) around the goal location.

For **Cabinet** tasks, agent needs to grasp the cabinet handle using its end effector, open the drawer, and keep pulling until it reaches a goal location. In PixMC, agent constantly receives a variable-valued reward inversely proportional to end effector's distance to cabinet handle. In PixMC-Sparse, agent only receives a fixed-valued reward when the mesh of its end effector intersects with the mesh of cabinet handle (visualized as the green sphere) around the goal location.

For **Pick** tasks, agent needs to grasp a cube-shaped object using its end effector and lift up the object until it reaches a goal height. In PixMC, agent constantly receives a variable-valued reward inversely proportional to end effector's distance to object, and a lift bonus when it starts to successfully lift up the object (i.e. when object is above the table). In PixMC-Sparse, agent does not receive any distance-to-object-based reward, i.e. no reward until it starts to lift up the object.

For **Move** tasks, agent needs to grasp a cube-shaped object using its end effector and lift up the object until it reaches a goal location. In PixMC, agent constantly receives a variable-valued reward inversely proportional to end effector's distance to object, and a lift bonus when it starts to successfully lift up the object (i.e. when object is above the table). In PixMC-Sparse, agent does not receive any distance-to-object-based reward, i.e. no reward until it starts to lift up the object.