Appendix

A. Results for Different Image Encoders

To evaluate whether our model provides improvements over other visual backbones, we train a version of ENTL that uses the Masked Vision Pre-training (MVP) [**?**] encoder pre-trained on embodied robotics data. We evaluate the model on the ObjectNav task, and as Table 1 shows the CLIP based model outperforms the MVP based one.

Model	# Frames	Success	SPL
ENTL w/ MVP [**?**]	20M	0.47	0.15
ENTL w/ CLIP [**?**]	20M	0.50	0.18

Table 1. **Encoder Ablation.**

B. ProcTHOR Data Collection

We collect 20 Million frames of demonstrations in the ProcTHOR environment which we use for training the joint ObejctNav and PointNav ETL model. Since there is no large scale human demonstration dataset in this environment, we collect trajectories guided by a planner with full information about the scene. The planner plots the optimal path from the current location of the agent to a valid goal state for a given object category. This planner is implemented in the Unity environment upon which AI2-THOR is built. A goal state is defined as any state in which the agent is within 1.5m of the target object, and the object is visible (as defined by the RoboTHOR ObjectNav challenge).

Since a direct trajectory from the current location to a goal does not capture any exploratory behavior, we instead structure our trajectories as tours of every target object in a given scene. Each trajectory is obtained by navigating the agent from a random starting location to a goal object, then from that object to every other goal object in the current scene in random order. This way the agent walks around an entire room in every trajectory.

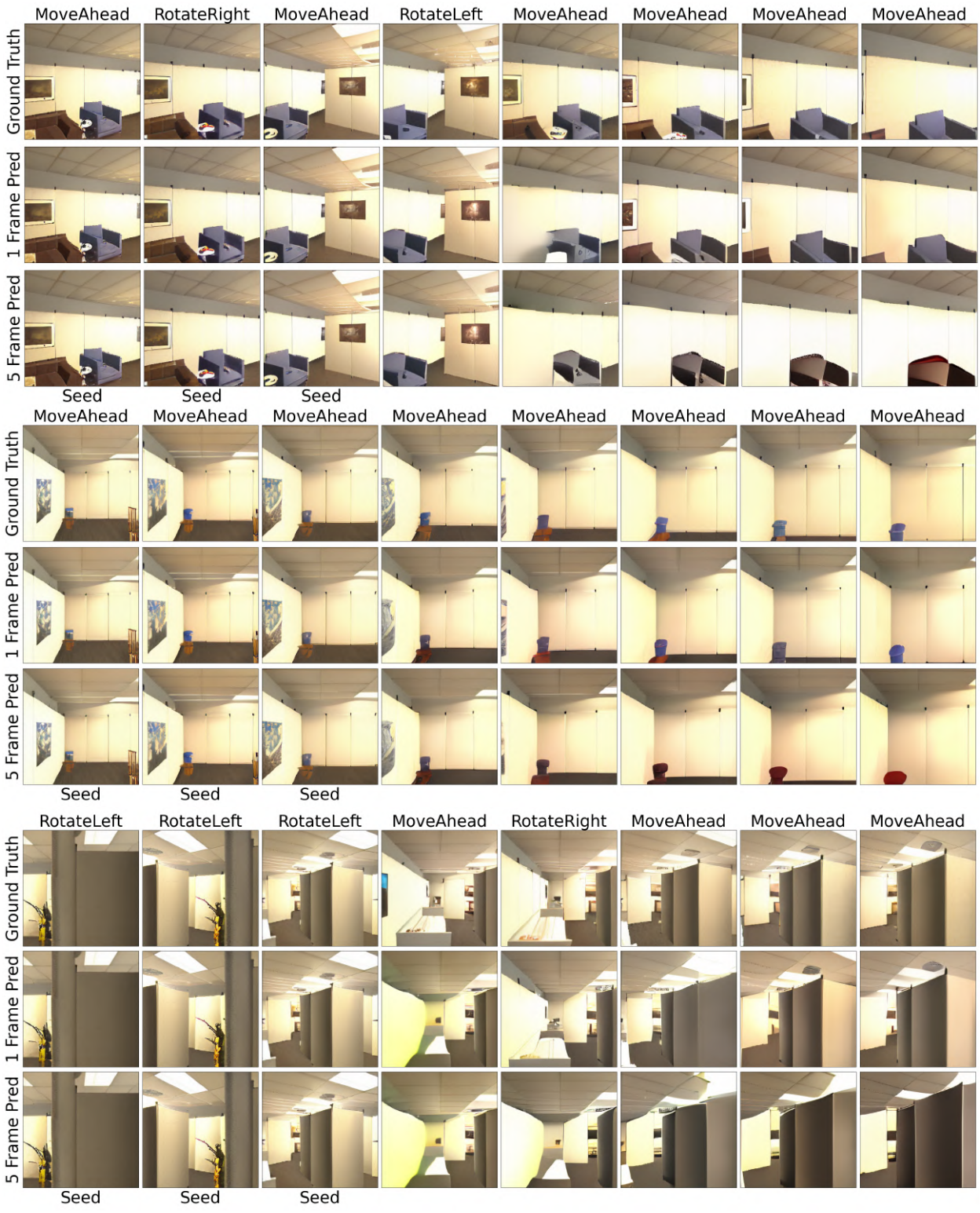
At each step we record the RGB frame, the action that was taken, the agent pose and the list of all ObjectNav target objects that are currently in range of the agent. Only RoboTHOR ObjectNav challenge goal objects are considered as goal objects.

Any state can also serve as a PointNav goal state, so we do not collect PointNav trajectories explicitly.

C. Additional Frame Prediction Examples

Additional rollouts produced by the ENTL model. As shown the model is able to predict fairly well if the room is empty or objects are given in the seed frames, but performs poorly when hallucinating new objects. Additionally, each consecutive predicted frame suffers from an accumulation of errors as the predicted image tokens have to be decoded into an RGB image by VQ-GAN, then passed through CLIP to predict the next frame. The results are shown in 1.

Figure 1. Additional predictions of future frames in unseen RoboTHOR scenes. For each rollout 3 seed frames are given to the model, then it predicts the next 5 steps with and without re-seeding after each step, based on the provided actions.




MoveAhead MoveAhead MoveAhead MoveAhead MoveAhead MoveAhead MoveAhead MoveAhead
Ground Truth
1 Frame Pred
5 Frame Pred
Seed Seed Seed
MoveAhead RotateRight MoveAhead RotateRight MoveAhead MoveAhead MoveAhead RotateLeft
Ground Truth
1 Frame Pred
5 Frame Pred
Seed Seed Seed
MoveAhead MoveAhead RotateRight MoveAhead RotateRight RotateRight MoveAhead MoveAhead
Ground Truth
1 Frame Pred
5 Frame Pred
Seed Seed Seed