Appendix for Pathdreamer: A World Model for Indoor Navigation

Jing Yu Koh[1] Honglak Lee[2] Yinfei Yang[1] Jason Baldridge[1] Peter Anderson[1]

[1]Google Research [2]University of Michigan

A. Qualitative Results

A.1. Val-Unseen Results

We present additional qualitative results generated by the Pathdreamer model, on the Val-Unseen split of the R2R dataset. These environments are not seen by the model during training, and act as a test of the generalization ability of Pathdreamer. Figure 1 presents cherry-picked examples of Pathdreamer generated sequences and Figure 2 presents randomly chosen examples.

A.2. Val-Seen Results

The Val-Seen split contains novel navigation trajectories for environments seen in the training split. As described in the main paper, the Structure Generator and Image Generator models perform significantly better on Val-Seen, since they can memorize the training environments with less generalization required. We observe that generation results maintain high fidelity even at large distances from the location of the input observation. Figure 3 presents cherry-picked examples and Figure 4 presents random examples.

A.3. Noise Interpolation

As described in the main paper, the Structure Generator is able to generate alternative, diverse, *room reveals* for a given scene. In areas without guidance pixels (i.e., in areas of the environment not seen from a previous view), interpolating over different noise vectors produces diverse and plausible outcomes. Figure 5 presents cherry-picked examples of Pathdreamer sequences when conditioned on different noise vectors. Figure 6 contains additional examples of randomly selected sequences conditioned on random noise vectors.

B. Generated Videos

In addition to image generation, Pathdreamer is capable of generating continuous video sequences, simply by sequencing generated images with small viewpoint changes. We provide videos displaying Pathdreamer generated results for several unseen environments. We also compare Pathdreamer's rendering quality using multiple observations to the output of the Habitat simulator using mesh-based rendering, illustrating a favorable comparison in quality. We refer readers to the YouTube link[1] for the video results.

C. Implementation Details

Structure Generator We trained this model with a batch size of 64, over 50 epochs. For the first 30 epochs the model is trained with teacher forcing, i.e., the previous observation that is used as input at each step is the ground-truth previous observation. During the teacher forcing stage, the number of ground truth context frames used for a path of length L is decayed uniformly from $L-1$ to 1. After 30 epochs, we switch to the recurrent setting in which the model's previous prediction is used as the previous observation at each step during the rollout (similar to our setup at inference time). We use the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We apply a learning rate which starts at $1e^{-4}$ and warms up to $2e^{-4}$ uniformly over 10 epochs.

Image Generator This model was trained with a batch size of 128 over 500 epochs. We use the Adam optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a learning rate of $2e^{-4}$ for both the generator and the discriminator. During training, the discriminator is trained for 2 steps for each generator training step. Following standard practice, at inference time we applied an exponential moving average (EMA) to the generator weights with 0.999 decay. For the choice of loss weights, we set $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{VGG}} = 0.07$, and $\lambda_{\text{FM}} = 0$. We found experimentally that excluding the feature matching loss speeds up training throughput and did not have a significant effect on the results.

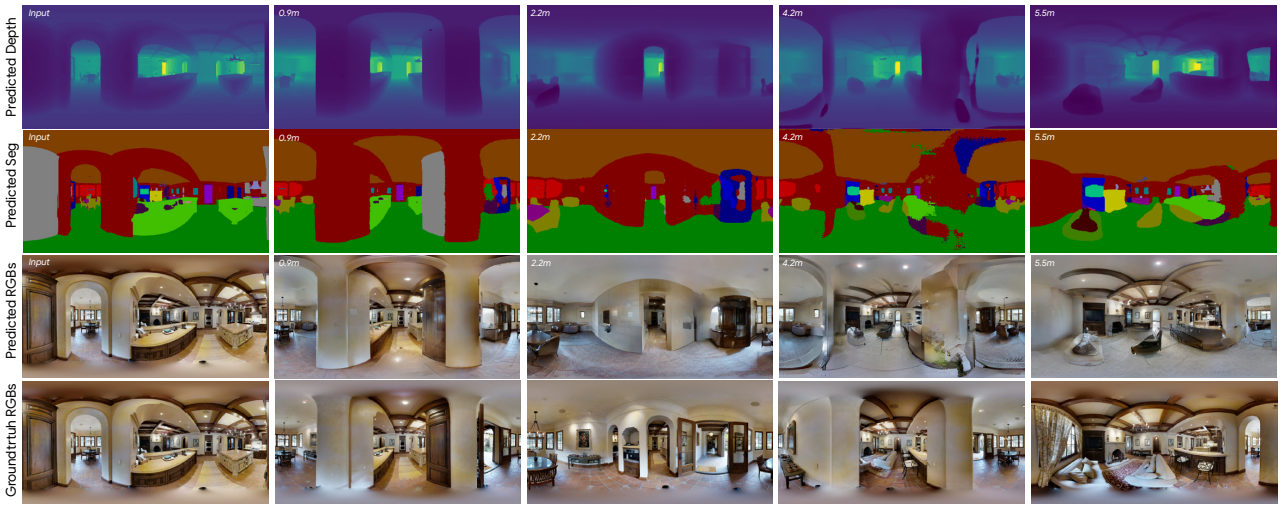
Evaluation Details We compute the FID score[2] using 10,000 random samples for each prediction sequence step. As the R2R validation sets contain 783 and 340 sequences for Val-Unseen and Val-Seen respectively, we perform data

[1] https://youtu.be/StklIENGqs0

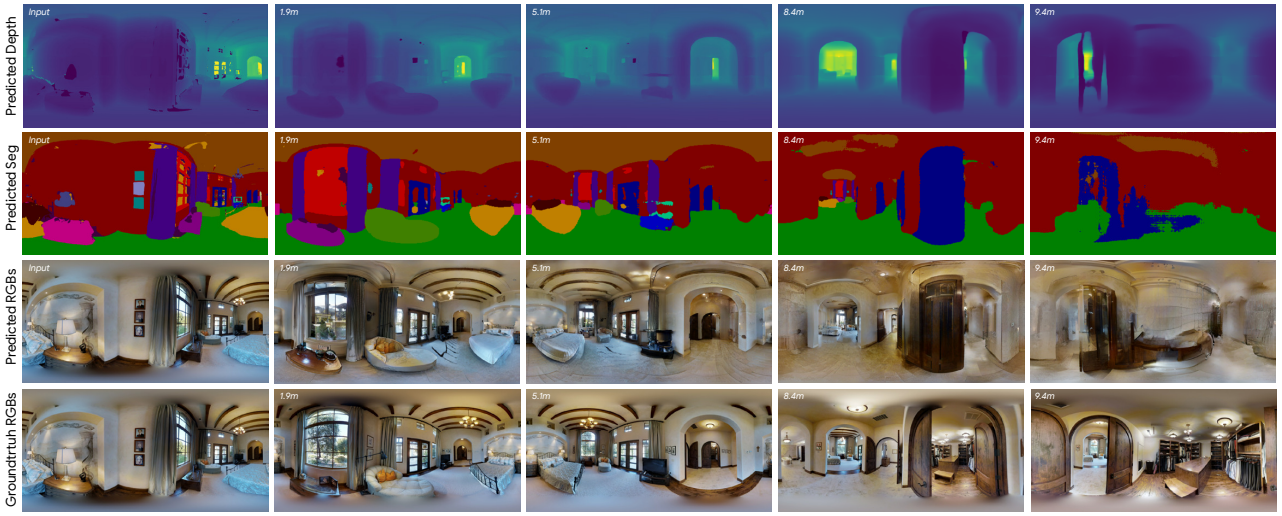
[2] We used https://github.com/mseitzer/pytorch-fid for computing results.

augmentation with random horizontal roll and flips to acquire 10,000 samples.

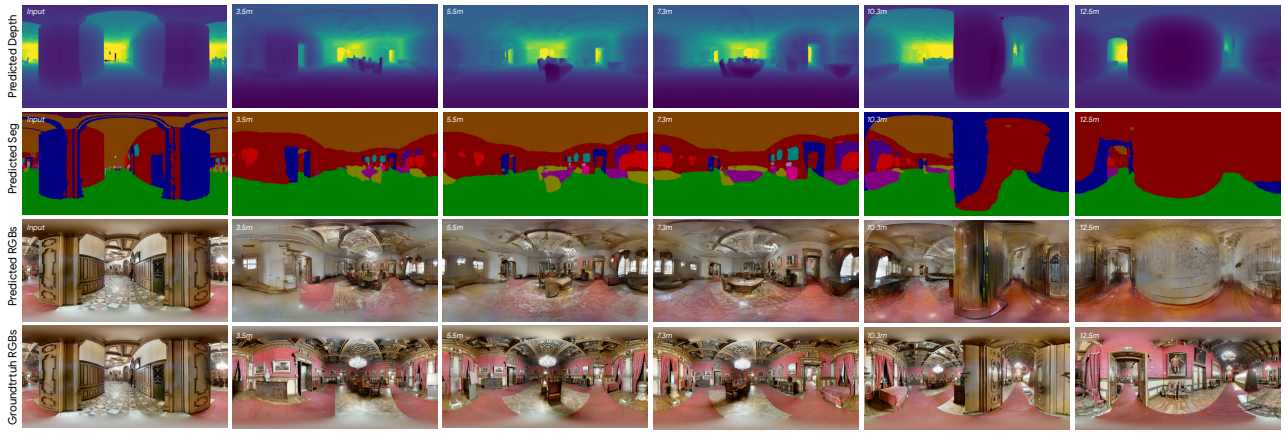
We run evaluation every 2000 training steps, and select the best checkpoint on the Val-Seen and Val-Unseen set for reporting results for the respective split. We note that training time is generally significantly longer on Val-Seen as compared to Val-Unseen. Due to the benefit of overfitting on the training set for Val-Seen, performance continues to improve on even as generalization performance on Val-Unseen deteriorates.



(a) At 5.5m, Pathdreamer completes a room reveal – imagining a living room space. Despite differences in details (e.g. door on the left rather than a window, and placement of a couch closer to the right), the result is quite close to the ground truth.

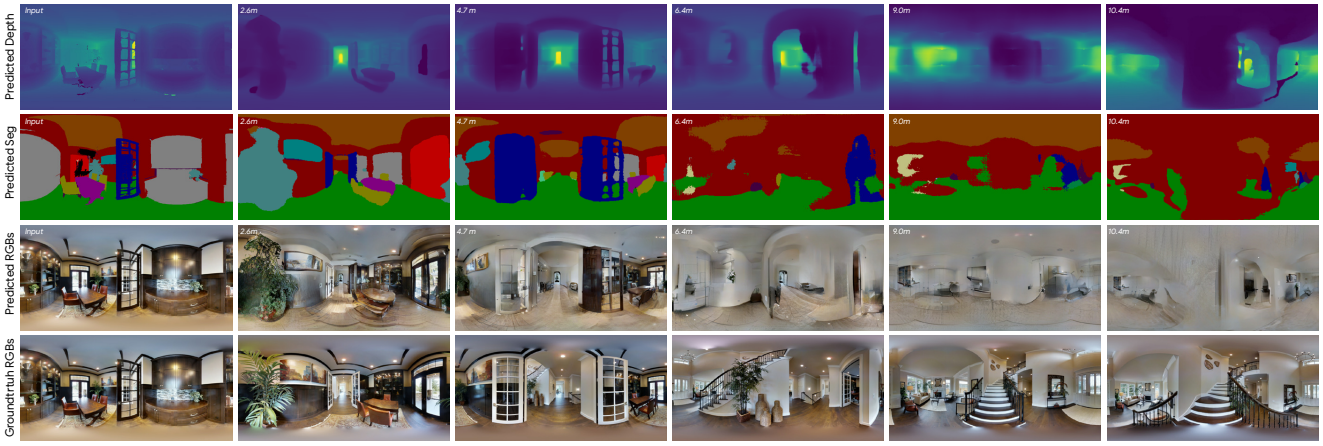


(b) In this example, the navigation sequence exits a bedroom. Pathdreamer is able to maintain consistency and realism for unseen corners of the room (1.9m and 5.1m). At 8.4m, Pathdreamer also generates a plausible hallway seen after exiting the room.

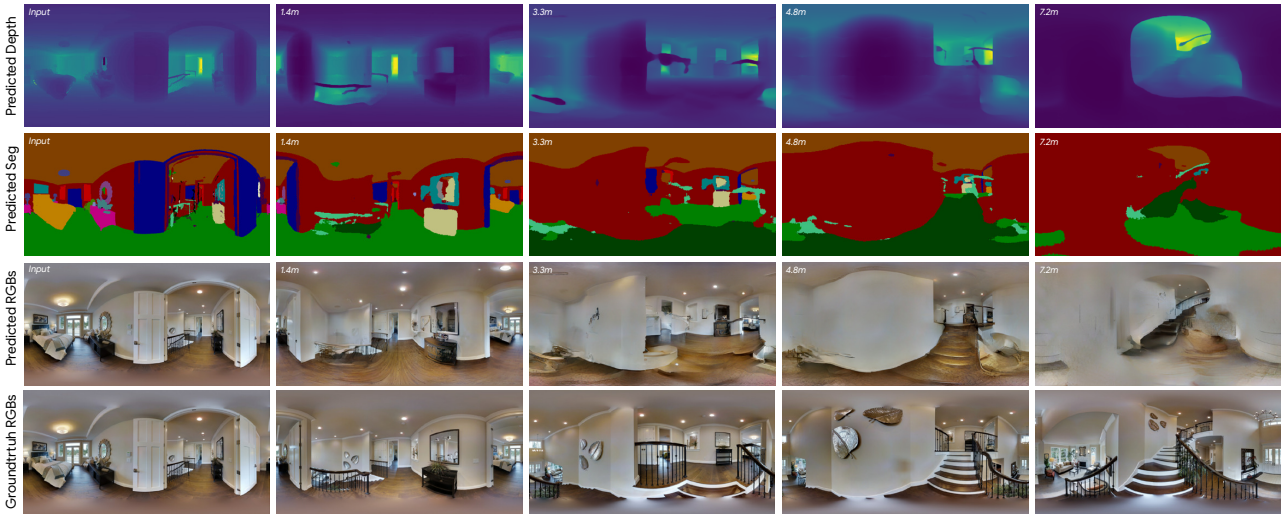


(c) Despite the initial provided observation being quite distant, Pathdreamer is able to synthesize realistic segmentation and RGB details of the table, paintings, and doorway at 3.5m after moving closer. A plausible hallway scene is also synthesized at 10.3m and 12.5m.

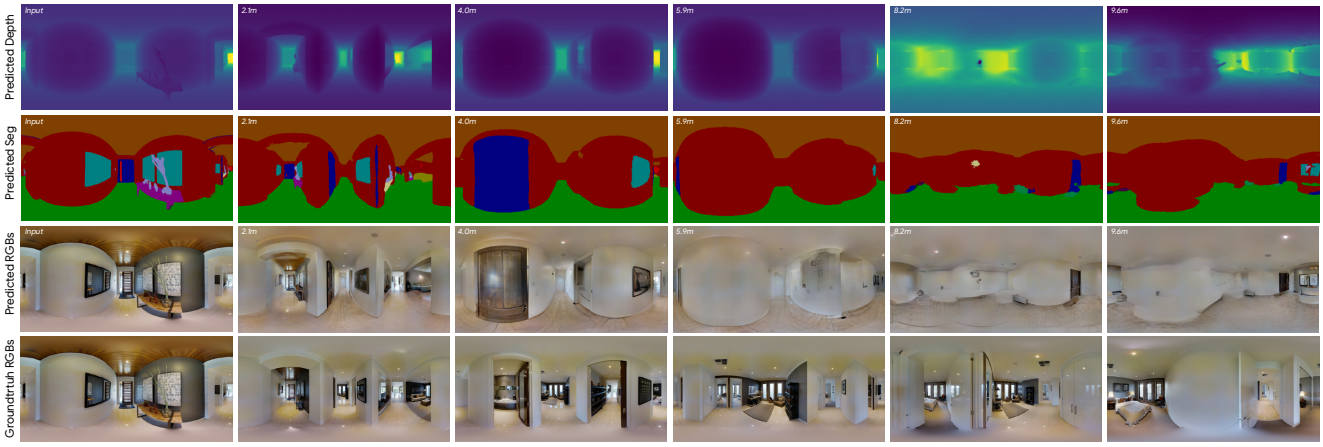
Figure 1: Selected examples of predicted sequences from the Val-Unseen split using one ground truth observation as context.



(a) In this sequence, feature predictions in the segmentation space are generally preserved up to 4.7m. The predictions tend to diverge from the groundtruth after, due to the failure of the model to predict the existence of a stairwell at 4.7m

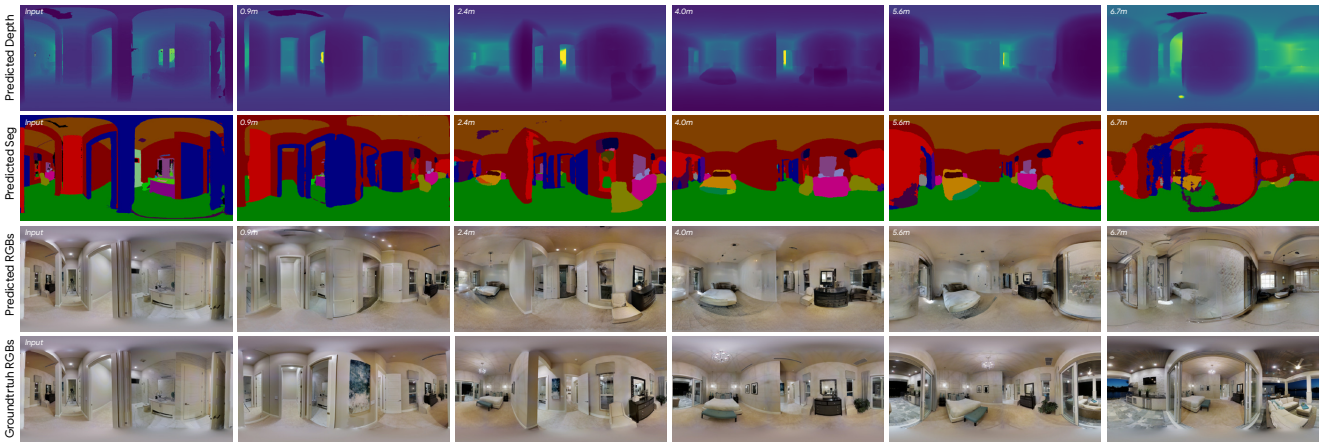


(b) Model predictions in the segmentation space are generally accurate up to 4.8m. However, the corresponding RGB outputs are not as realistic, likely due to difficulties in generating intricate objects such as stair railings.

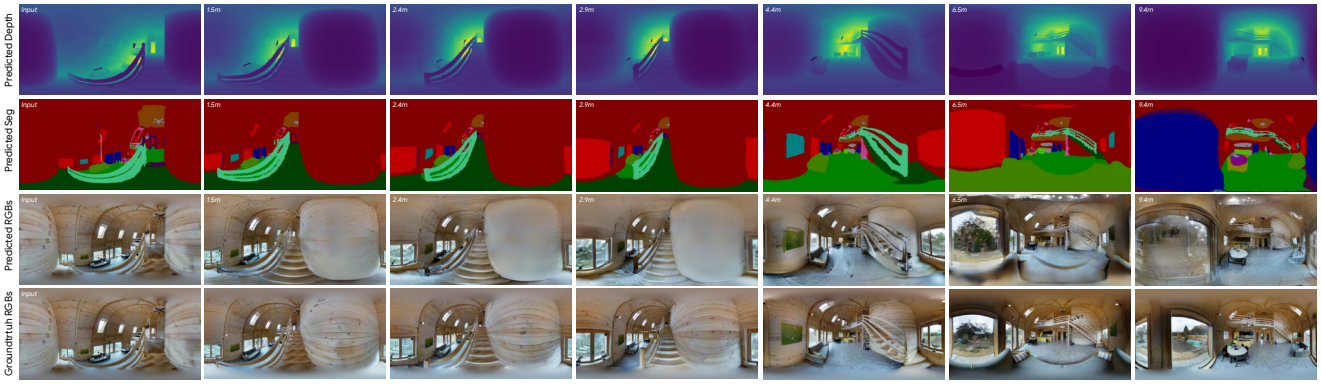


(c) In this example, the model fails to predict the existence of the room entrance in the center of the panorama at 2.1m, which leads to the outputs eventually diverging from the ground truth. Despite this, predicted results remain generally plausible.

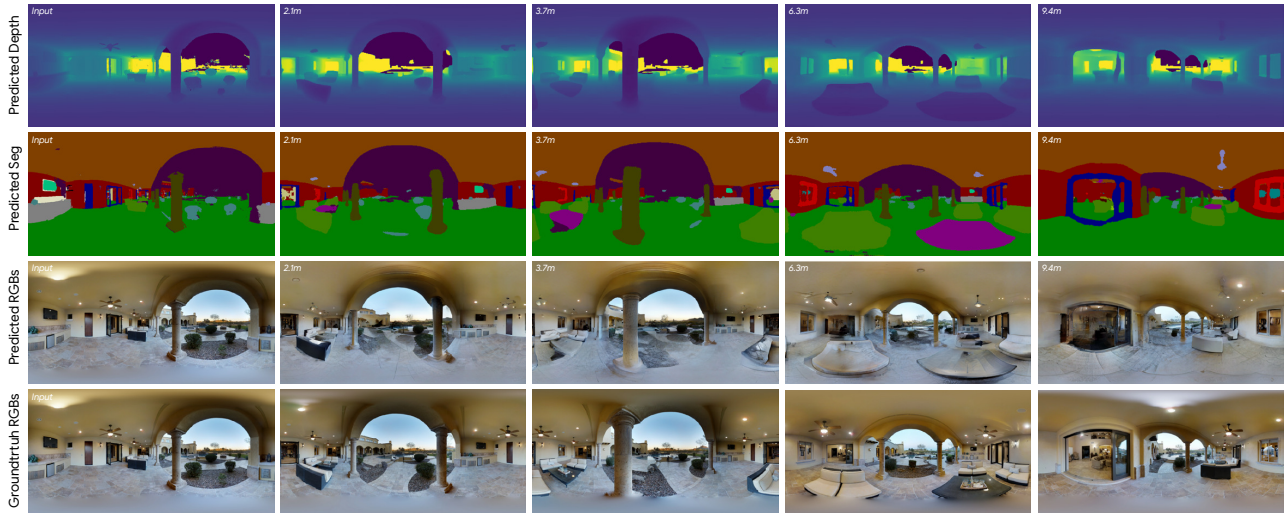
Figure 2: Randomly selected prediction sequences from the Val-Unseen split using one ground truth observation as context.



(a) Pathdreamer is able to almost perfectly recreate the navigation sequence, despite only being provided with a single ground truth observation. Several details are missed in the RGB outputs, such as the sliding glass doors at 5.6m. This is likely due to the poor quality depth returns from glass doors in the training data.

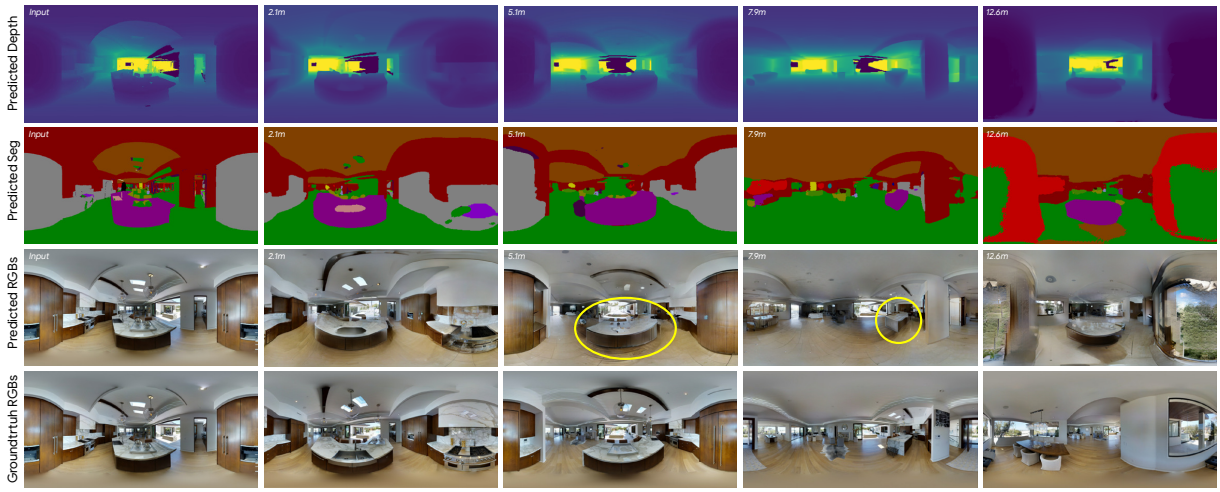


(b) From the initial observation, Pathdreamer is able to recreate the scene from multiple perspectives: (1) the middle of the stairs at 1.5m, (2) the bottom of the stairs at 4.4m, and (3) from the corner of the room at 6.5m. The ability to accurately representing geometry is one of its strengths as a world model.

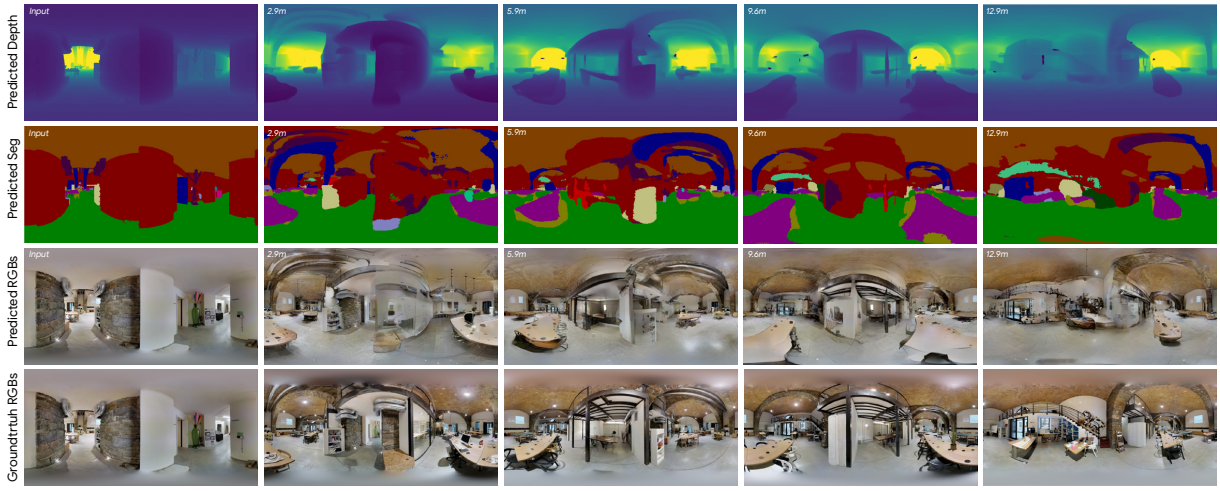


(c) In this example, the outdoor scenery is accurately recreated due to the information from both the segmentation predictions and the projected RGB values. Some textural details are lost in the couch and table at 6.3m, which suggest the potential to improve results by training a stronger Image Generator.

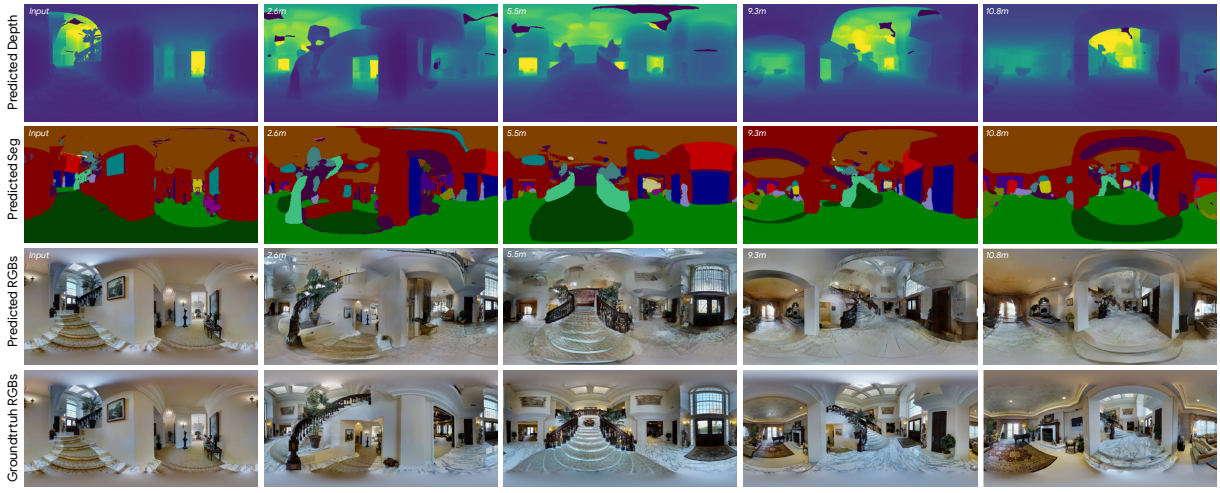
Figure 3: Selected examples of predicted sequences from the Val-Seen split using one ground truth observation as context.



(a) Pathdreamer is able to perform view synthesis for potentially unbounded viewpoint changes. Provided with an observation of the kitchen counter from one angle, realistic views can be synthesized from other perspectives (highlighted at 5.1m and 7.9m).

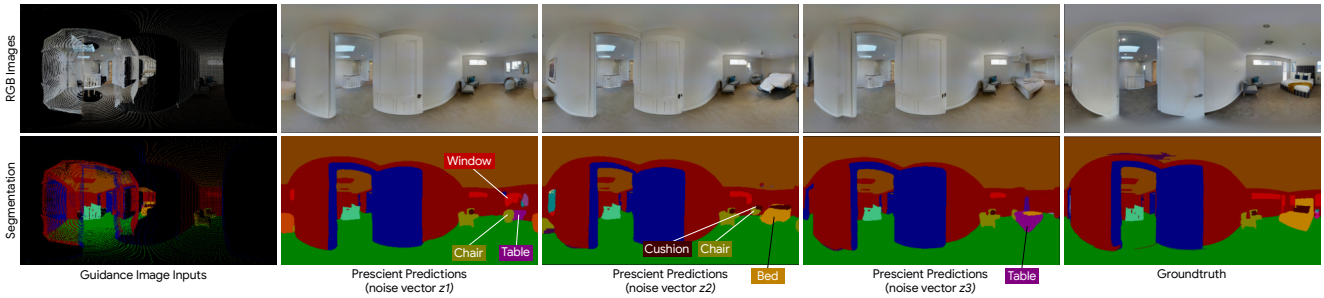


(b) While Pathdreamer is able to maintain consistency of the generated semantic segmentations in this example, the complexity of the structures (e.g. arched ceiling beams) make RGB generation difficult, and the results are blurred in certain regions.

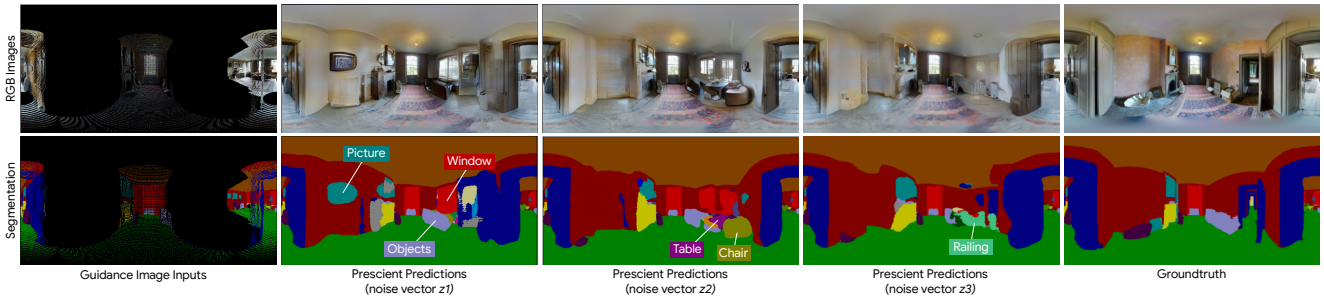


(c) Pathdreamer is able to handle complex scene geometry, as shown in this example with multiple floors. While the overall scene looks fairly realistic, the Image Generator has some trouble with generating realistic looking stairs (e.g. at 5.5m).

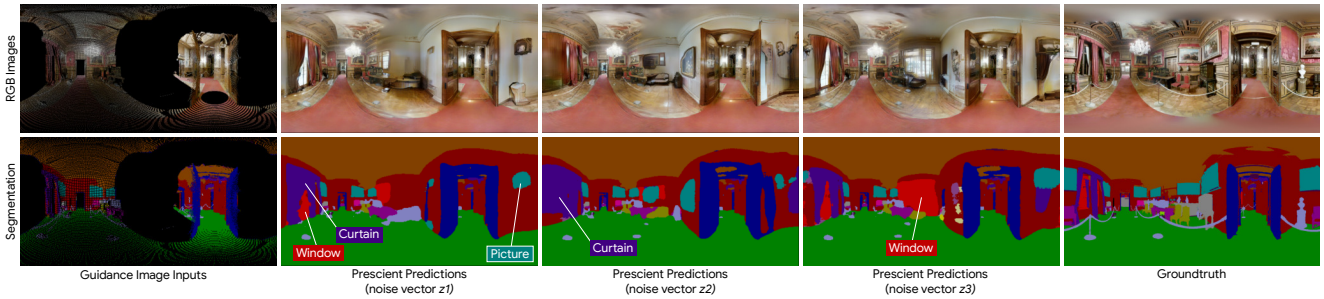
Figure 4: Random prediction sequences from the Val-Seen split using one ground truth observation as context.



(a) Several plausible scene layouts are predicted for the unknown region of the image on the right (indicated in black in the guidance inputs). Notably, Pathdreamer is able to synthesize a complete set of a bed, cushion, and chair in the second example (noise vector z_2).



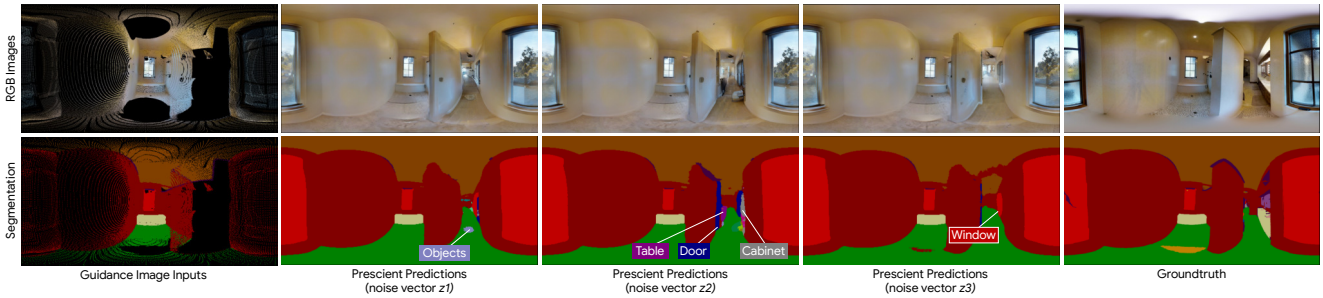
(b) Pathdreamer predicts several plausible scenes: (1) a room with a painting on the left wall, (2) a room without a painting, but with a table and chair on the right, and (3) a room without a table or chair, but with stairs leading down.



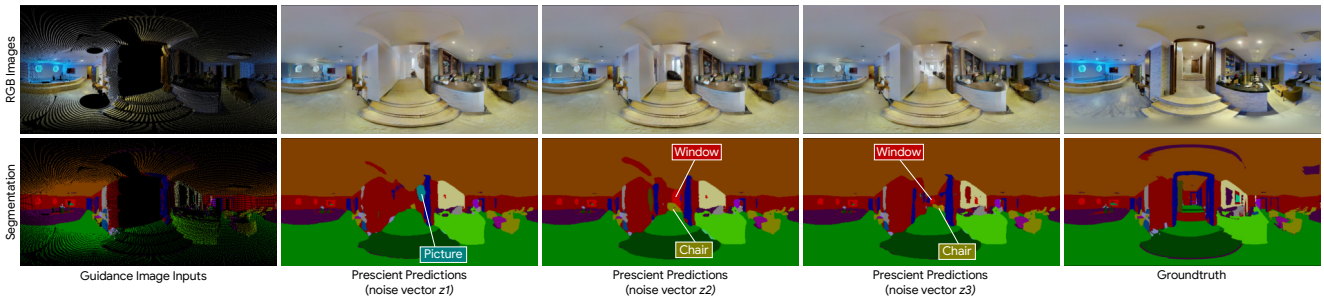
(c) Several varying but plausible scenes are considered by Pathdreamer: (1) a window with curtains open at the side, (2) a window covered by curtains, and (3) a large floor to ceiling window on the right of the room.

Figure 5: Diverse and semantically plausible predictions can be sampled for scenes containing previously unseen regions. Shown here are several selected examples, each consisting of three alternative room reveals and the corresponding ground truth (from Val-Unseen).

(a) A wall is correctly predicted to exist on the right of the image. Varying layouts and decorations for the wall are proposed, which all create valid segmentation and RGB generation results.



(b) In this example, the missing region of the image is correctly predicted to be a hallway. Several plausible variations of the hallway are synthesized, each containing objects that are likely to be present.



(c) This example contains stairs leading to a hallway, which is correctly predicted by the model. Several potential layouts are synthesized, although none of them fully synthesize the entire door frame present in the ground truth example.

Figure 6: Random examples of generation results when sampled with different noise vectors.