7 Appendix

7.1 Dataset Detail

All the data are collected at a rate of 20Hz, both state and visual inputs. State information includes 6 dof torso localization in the global frame, leg joint angles (left, right calf and thigh), torso velocity, torso angular velocity, and average gait frequency in 2 seconds moving window. Visual information includes RGB images, an aligned depth frame from D455 at a resolution of 848x480, semantic segmentation masks by DINOv2 + Mask2Former segmentation head, as well as panoramas with all the aforementioned channels projected to 360 view. This brings the total dataset to 198 minutes, and over 400GB.

We optimize storage by retaining depth frames only when there's a significant change in camera position—specifically, an angular movement exceeding 15 degrees or a translation beyond 1 meter. This strategy reduces the point cloud processing rate to approximately 1 keyframe per second at normal walking speeds, thus conserving computational resources for semantic segmentation. Visual memories are generated in a separate process at a 20Hz, aligned with the current camera pose and buffered depth frames, to support timely model inference.

The raw data from stereo cameras often requires prepossessing to enhance quality. We apply a tuned Canny edge filter to mitigate artifacts along object edges, removing up to 10 pixels around the disjoint edges, and preparing the depth frames for more accurate panorama construction.

We define eight semantic classes for this purpose: No label, normal ground, stair, door, wall, obstacle, movable, and rough ground. The collision evaluation framework specifically considers ground, stairs, walls, obstacles, and rough terrains to accurately differentiate between navigable and non-navigable spaces.

7.2 Diffusion Model Detail

Fig. 9 illustrates the detailed model architecture. We use the transformer encoder blocks which contains multi-head self-attention layers. They are stacked between the Down and Up blocks of the UNet. The conditions of the generation (past trajectory and most recent visual memory) are first turned into embeddings, and then passed into the Down and Up blocks.

In training, instead of cutting up the dataset into unique sub-trajectories, we take overlapping sub-trajectories to further augment the dataset. This way, there is a total of more than 220,000 diverse sub-trajectories for the model to train on.

At inference time, one can either use DDIM, DDPM or our hybrid inference method. Specifically the hybrid inference involves first getting a rough convergence with 20 steps of DDIM, the time steps are sub-sampled uniformly from the total steps. And then the last 5 steps are performed with DDPM step 5 to 0 to bring the sample to full convergence. The variance we used in DDPM is fixed small, or mathematically:

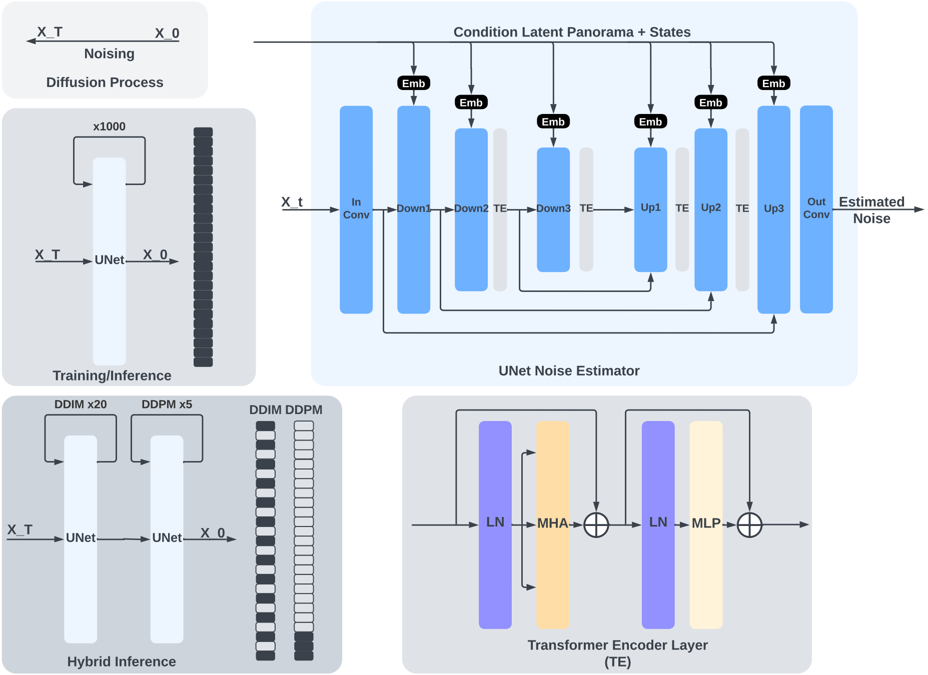

Fig. 9: Diffusion Model: Architecture and hybrid generation details

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} z_\theta(x_t, t)) + \sigma_t z \tag{3}$$

Where:

$$\sigma_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{4}$$

7.3 Collision-Free Metric Detail

To calculate the collision-free metric, we first re-project the visual memory provided as prediction condition into a 3D point cloud. At each predicted time step, we retrieve the closest 20 points in the point cloud by K-D Tree. If there are more than 10 points within the 16cm radius circle, the current and all subsequent time steps are deemed collided. The collision metric value will then be the index collided time step. Therefore the maximum score is n-1 if prediction contains n steps. The parameters are tunable, and these are the empirical values we found to be most reasonable.