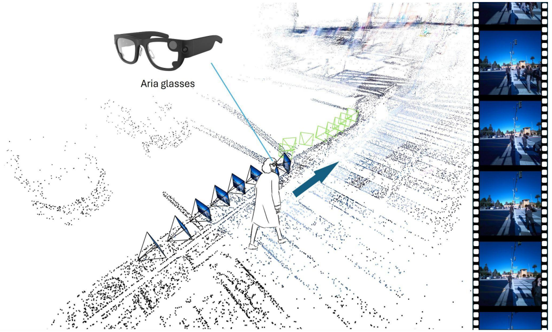
LookOut: Real-World Humanoid Egocentric Navigation

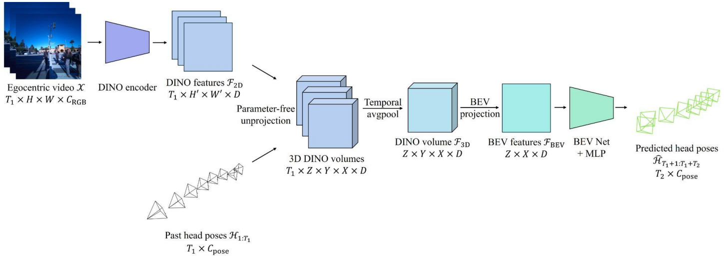
Boxiao Pan, Adam W. Harley, C. Karen Liu*, Leonidas J. Guibas*
bxpan@stanford.edu Stanford University * Equal advising

Overview



Given a posed egocentric video (black-outlined frustums, with frames shown in detail on the right), our model predicts a sequence of 6D head poses in the future (green-outlined frustums). We design a data collection pipeline with the Project Aria glasses and train our model on a dataset collected this way. This problem features real-world navigation challenges including collision avoidance with static and dynamic obstacles, and human-like information-gathering behaviors (e.g. looking to the sides when crossing roads in this example). The point cloud is shown for visualization but is not an input to the model.

LookOut Architecture



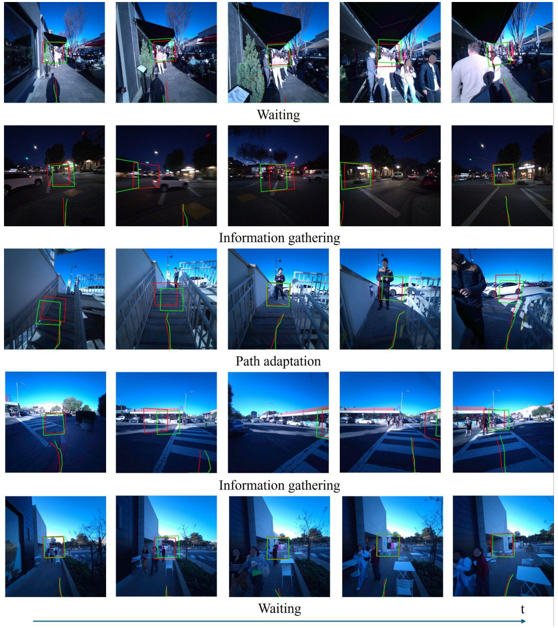
Given a posed egocentric video, we obtain frame-wise DINO features with the pre-trained encoder, and unproject them to 3D for temporal aggregation. The aggregated features are then projected to BEV for further processing and eventually used to predict future head poses. The bulk of the computation happens in BEV through BEV Net (ConvNets + MLPs).

Aria Navigation Dataset (AND)

We need data that 1) contains posed egocentric RGB videos of real-world navigation with human, 2) captures both static and dynamic obstacles, and 3) displays the active information-gathering behaviors that we want our model to learn. Additionally, we would like our capture setup easily scalable.

We designed a data collection pipeline that uses only a pair of the Project Aria glasses as the hardware, and requires a few seconds to setup before each recording session. Our resulting dataset contains 4 hours of recording from 18 densely populated places.

Qualitative Evaluation



Red: model predictions; Green: gourd-truth.
Box: viewing frustums; Curve: ground-projected translation.

- Our model forecasts collision-free paths around both static and dynamic obstacles.
- Our model learns the information-gathering behavior that humans demonstrate in the training data.
- Other interesting behaviors emerge, such as waiting when no paths available, and path adaptation based on new visual cues.

Quantitative Evaluation

Baselines	L_1_trans ↓	L_1_rot ↓	Col_stt_avg ↑	Col_dyn_avg ↑
Linear Extrapolation	0.45	1.21	79.1	82.4
EgoCast [8]	0.34	0.63	85.3	86.2
Ours	**0.17**	**0.16**	**85.6**	**90.2**

L_1: L1 error on translation / rotation.
Col_*_avg: percentage of predictions that are at least x (cm) away from the closest obstacle, averaged between x ∈ {15, 25, 35}. stt: static; dyn: dynamic.

Ablations	L_1_trans ↓	L_1_rot ↓	Col_stt_avg ↑	Col_dyn_avg ↑
Point Cloud Only	0.40	0.88	83.2	84.6
Depth Only	0.22	0.23	87.0	91.6
RGB + Depth	0.15	0.13	87.4	91.4
DINO temp pooling	0.26	0.44	84.9	86.2
3D Conv	0.17	0.19	85.6	89.9