LookOut: Real-World Humanoid Egocentric Navigation

Boxiao Pan   Adam W. Harley   Francis Engelmann   C. Karen Liu[*]   Leonidas J. Guibas[*]

Stanford University

Abstract

The ability to predict collision-free future trajectories from egocentric observations is crucial in applications such as humanoid robotics, VR / AR, and assistive navigation. In this work, we introduce the challenging problem of predicting a sequence of future 6D head poses from an egocentric video. In particular, we predict both head translations and rotations to learn the active information-gathering behavior expressed through head-turning events. To solve this task, we propose a framework that reasons over temporally aggregated 3D latent features, which models the geometric and semantic constraints for both the static and dynamic parts of the environment. Motivated by the lack of training data in this space, we further contribute a data collection pipeline using the Project Aria glasses, and present a dataset collected through this approach. Our dataset, dubbed Aria Navigation Dataset (AND), consists of 4 hours of recording of users navigating in real-world scenarios. It includes diverse situations and navigation behaviors, providing a valuable resource for learning real-world egocentric navigation policies. Extensive experiments show that our model learns human-like navigation behaviors such as waiting / slowing down, rerouting, and looking around for traffic while generalizing to unseen environments. Check out our project webpage at `https://sites.google.com/stanford.edu/lookout`.

1. Introduction

Navigating safely in the real world from egocentric observations is an ability that we humans possess, yet extremely difficult for machines to learn. This is largely due to the diverse and complex situations that exist in practical scenarios. Such capabilities are crucial for various applications including humanoid robotics [42], VR / AR [8], and assistive navigation [63].

Many works have approached this problem from different angles. Vision-Language Navigation (VLN) [18, 21, 25, 39, 66] focuses on localizing goals and planning long-term goal-directed paths, typically in simulated static environments. Robotic social navigation [14, 17, 36, 41, 56] learns socially compliant navigation policies in dynamic environments. These works generally target wheeled or legged navigation robots, whose action and observation distributions are vastly different from the human form factor. Recently, several works investigate egocentric navigation that predicts trajectories [63] or full-body poses [8] for humans. They however assume static environments.

Despite these advances, a real-world deployable humanoid egocentric navigation policy remains challenging. First, a method for humanoid navigation in dynamic environments is lacking. Second, existing methods ignore an important aspect of human-like navigation, which is the active information gathering via head turning. Humans often rotate their heads and look for useful information. For example, we look to the side before crossing roads to check any passing vehicles, look downward when stepping off / onto curbs, etc. This ability is helpful for real-world deployment, partly due to the limited FoV of cameras. Lastly, we do not have a way to collect multi-modal labeled training data at scale due to the difficulty of deploying humanoid robots in the real world.

In this work, we make steps towards a real-world deployable humanoid navigation policy from all three of these fronts. To tackle the challenge of 3D dynamic scene awareness, we propose a model that unprojects per-frame DINO [2, 38] features to 3D, and aggregates the 3D feature volumes across time to gain a holistic understanding of the geometric constraints posed by the environment. Moreover, through training on our collected dataset that contains extensive dynamic obstacles (*i.e.* pedestrians and vehicles), our model effectively learns the ability to navigate around both static and dynamic objects. To model active information gathering, we design our framework to predict 3D head rotations in addition to translations (*i.e.* 6D head pose prediction), which can be used to calculate velocity commands normally input to humanoid robots [4, 46, 47]. Additionally, in our data collection process, we ask our human subjects to follow a careful information-gathering strategy, *e.g.* always looking for passing cars before crossing roads. To

[*] Equal advising

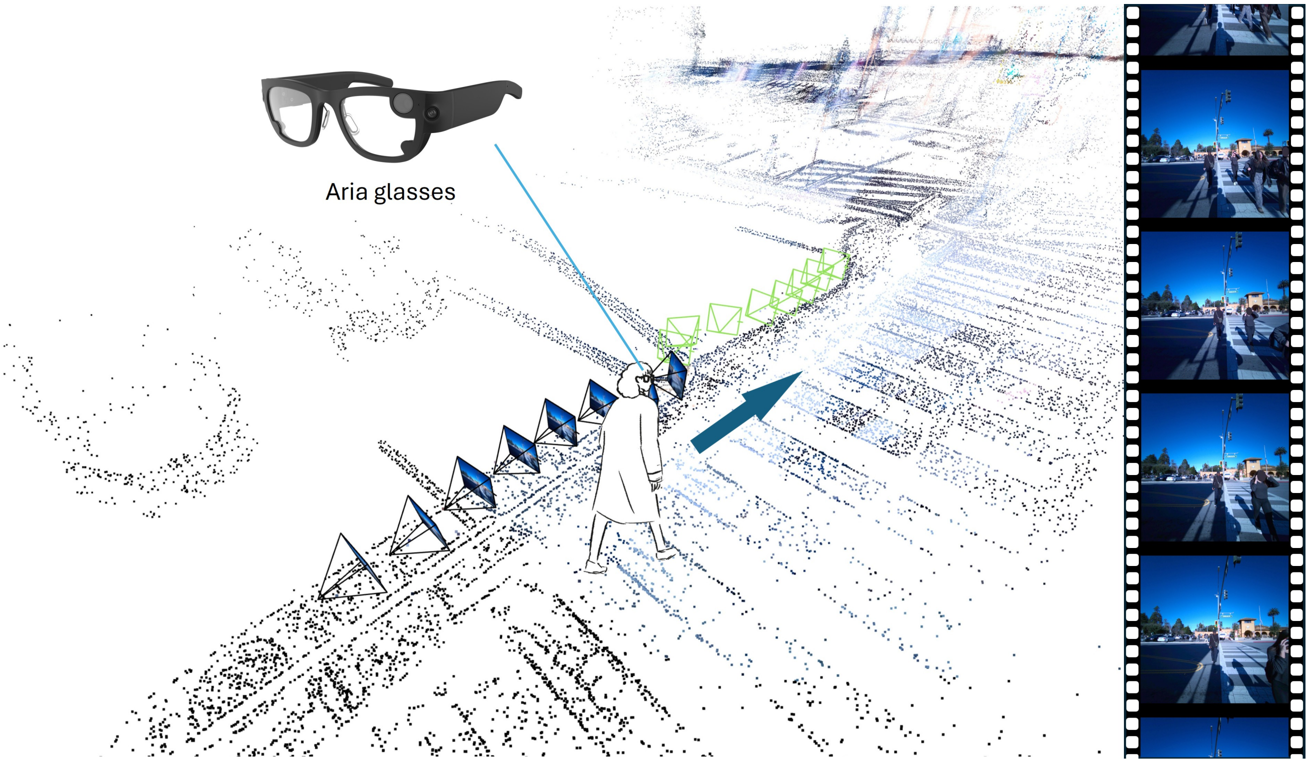

Figure 1. **Problem formulation**. Given a posed egocentric video (black-outlined frustums, with frames shown in detail on the right), our model predicts a sequence of 6D head poses in the future (green-outlined frustums). We design a data collection pipeline with the Project Aria glasses and train our 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.

solve the challenge of collecting useful data at scale, we propose a data collection pipeline that uses a pair of Project Aria glasses [7] as the data collection tool. This pipeline enables naturalistic human navigation demonstration collection without drawing attention [42]. Unlike traditional collection pipelines that require carefully mounting various sensors [42] or teleoperating robots [17], our pipeline is extremely easy to set up, taking only a few seconds at the beginning of each recording session, while providing various data modalities including RGB videos, audio, eye gaze, SLAM reconstructed head poses and point cloud. It thus provides a way to scale up the data collection process with minimal effort. With this pipeline, we collected a dataset that consists of 4 hours of real-world navigation sessions, covering 18 densely populated places.

In summary, we make the following contributions: (1) we introduce the challenging task of 6D head pose trajectory prediction from posed egocentric observations, under the presence of static and dynamic obstacles. (2) We propose a model dubbed LookOut that aggregates unprojected 3D DINO features over time for semantic and geometric understanding, which proves effective in solving the task. (3) We contribute a data collection pipeline that leverages a pair of Project Aria glasses and requires minimal effort, providing a way to scale up the data collection effort at ease. (4) We collected a dataset with this pipeline, which consists of 4 hours of real-world navigation sessions and covers 18 places with dense and diverse traffic.

2. Related Work

Vision-language navigation. Studies on the task of Vision-Language Navigation (VLN) are arguably the most prevalent in the vision community around embodied navigation. This task is defined roughly as navigating to a specified goal location. Depending on the goal specification, the task has several variants as Point-Goal Navigation (PointNav) [68], Object-Goal Navigation (ObjectNav) [9, 21, 66], Image-Goal Navigation (ImageNav) [20], Language-Goal Navigation (LangNav) [39, 45], and Audio-Visual Navigation [3]. Some works also propose frameworks that unify several specifications [18, 25]. These works are generally developed in simulated environments, and focus on long-term path planning where no or simple [25] dynamic obstacles exist. In this work, we focus on short-term navigation whose main objective is collision-free locomotion.

Robotic social navigation. Classical methods have extensively studied goal-conditioned path planning for robots [52] that develop mathematical solutions given almost perfect knowledge on the environment. More relevant to our work are works that study egocentric navigation for robots. Robotic social navigation aims to predict a collision-free path [34, 44, 56] or higher-level instructions [32, 41] for robots, given egocentric sensing input such as RGB, LiDAR, and odometry. Common datasets for this task are collected by teleoperating robots [17] or human collectors wearing a sensor suite [36]. These works generally target wheeled or legged navigation robots. The actions and observations of such robots are drastically different from those of humanoids, due to their smaller scale, faster speed, and different morphologies. On the other hand, studies for humanoid robots [22, 23, 31, 37, 51] generally design classical methods with laser and stereo vision input for simplified environments.

Egocentric navigation for humans. Significant progress has been made in egocentric human motion estimation [11, 16, 26, 35, 50, 59, 65, 67], while forecasting future motion or trajectories that takes environment constraints into account is much less explored. COPILOT [40] predicts human-environment collision from multi-view egocentric videos in the form of collision labels and heatmaps. EgoNav [63] uses a diffusion model to forecast future trajectories from a chest-mounted RGBD camera input and past trajectories. EgoCast [8] predicts future full-body poses from a head-mounted RGB camera input and past head trajectories. Notably, EgoNav focuses on navigation while EgoCast studies diverse social and skilled human activities. However, all three works assume static environments, and do not learn the active information-gathering behavior critical in real-world scenarios. Moreover, we contribute a data pipeline that can be easily deployed at scale.

Egocentric datasets for humans. To study this task, we need a dataset that records real-world human navigation scenarios with significant presence of both static and dynamic obstacles, and provides data modalities on egocentric RGB videos, 6D head poses, and preferably also scene point cloud for collision checking. Traditional egocentric video datasets [6, 10, 27, 53] are generally captured in the monocular setting, and thus do not have camera pose annotations. The activities in these datasets are also diverse and do not focus on navigation. Some works [24, 40] propose synthetic data generation pipelines to simulate virtual humans walking in synthetic scenes. However, the generated motions are simple and unnatural due to the limitation of human motion generation methods. Autonomous driving dataset such as Waymo Open [58] comprehensively covers real-world traffic scenarios and provide dense 3D tracking for pedestrians and vehicles, but they do not record egocentric data for pedestrians. Recently, the Project Aria glasses [7] emerged as a convenient and natural way to record egocentric data. In particular, the Aria Machine Perception Service (MPS) provides an easy and highly optimized method to obtain accurate 6D camera (head) pose trajectories, environment point clouds, eye gazes, and more. Meta released several datasets collected with the Project Aria, and among them Aria Everyday Activities (AEA) [29] and Nymeria [30] record diverse indoor and outdoor activities. However, they either feature only single-human activities [29] or have multiple actors only in collaborative activities [30], so real-world navigation scenarios with potentially colliding agents are not captured.

3. Method

3.1. Problem Formulation

We illustrate our problem formulation in Fig. 1. Given a posed egocentric video $\mathcal{X} \in \mathbb{R}^{T_1 \times H \times W \times 3}$ and $\mathcal{H}_{1:T_1} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{T_1}\} \in \mathbb{R}^{T_1 \times 9}$, our goal is to predict the 6D head pose sequence for a short period in the future, *i.e.* $\mathcal{H}_{T_1+1:T_1+T_2} = \{\hat{\mathbf{h}}_{T_1+1}, \hat{\mathbf{h}}_{T_1+2}, \cdots, \hat{\mathbf{h}}_{T_1+T_2}\} \in \mathbb{R}^{T_2 \times 9}$. The head poses, which are also the camera poses, are parameterized as $\mathbf{h}_t = [\mathbf{t}_t | \mathbf{r}_t]$, where the rotation component $\mathbf{r}_t$ adopts the 6D continuous rotation representation [69]. Throughout our experiments, $T_1 = T_2 = 8$. The head poses are defined in a head-centered canonical frame specified in Sec. 3.3.

3.2. Model

The core functionality our model needs to have is extracting semantic and geometric information for the surrounding environment from a single monocular video stream. This prompts us to consider the following questions: (1) a strong visual encoder for semantic modeling, and (2) a way to aggregate information and reason in the 3D space. These motivate our key design choices of the model. For (1), we use the pre-trained DINO [2, 38] encoder to extract per-frame feature maps, due to its strong open-vocabulary semantic encoding capabilities [19, 60, 61, 64]. For (2), we adopt a strategy used in a number of object and scene representation models named "Parameter-Free Unprojection" [5, 12, 13, 54, 62]. Specifically, we bilinearly sample a subpixel 2D DINO features for each 3D coordinate defined in the canonical frame to obtain a 3D DINO feature volume. We then temporally aggregate the volumes for all time steps. These design choices endow our model with strong semantic and geometric reasoning capability while not having to rely on explicit geometric sensing input, such as depth and LiDAR. We illustrate our model in Fig. 2 and describe each component in detail next.

DINO feature encoding. We use the pre-trained DINO2 [38] variant `dinov2_vits14_reg` *, and apply it to each input frame (down-sampled to 224 × 224 spatial

*https://github.com/facebookresearch/dinov2

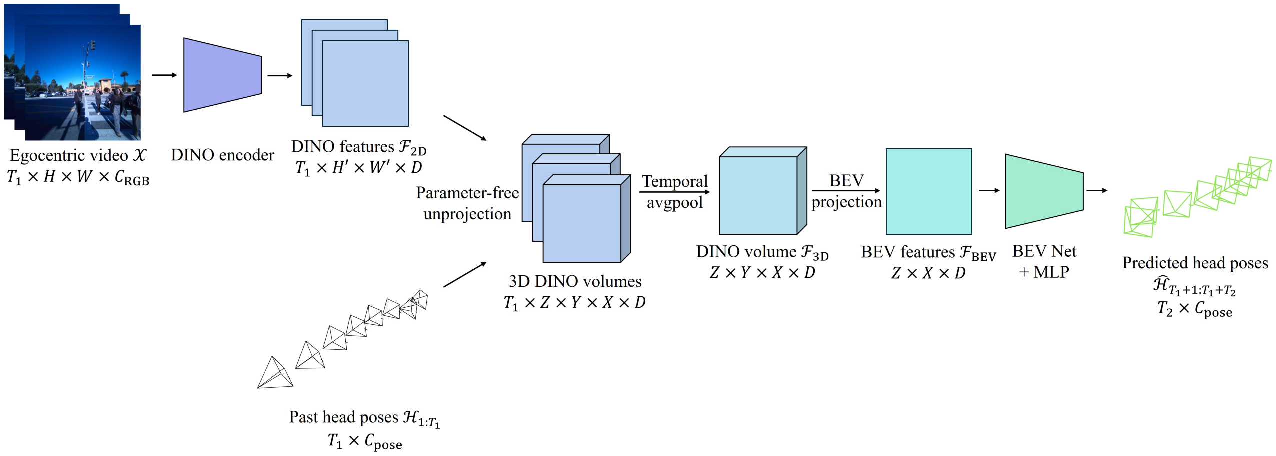
Egocentric video $\mathcal{X}$
$T_1 \times H \times W \times C_{\text{RGB}}$
DINO encoder
DINO features $\mathcal{F}_{\text{2D}}$
$T_1 \times H' \times W' \times D$
Parameter-free
unprojection
3D DINO volumes
$T_1 \times Z \times Y \times X \times D$
Temporal
avgpool
DINO volume $\mathcal{F}_{\text{3D}}$
$Z \times Y \times X \times D$
BEV
projection
BEV features $\mathcal{F}_{\text{BEV}}$
$Z \times X \times D$
BEV Net
+ MLP
Predicted head poses
$\hat{\mathcal{H}}_{T_1+1:T_1+T_2}$
$T_2 \times C_{\text{pose}}$
Past head poses $\mathcal{H}_{1:T_1}$
$T_1 \times C_{\text{pose}}$

Figure 2. **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.

resolution). This gives us a temporal sequence of 2D DINO features $\mathcal{F}_{\text{2D}} \in \mathbb{R}^{T_1 \times 16 \times 16 \times 384}$.

Parameter-free unprojection. Following [12, 13], we first define a voxel grid of 3D points in the canonical frame, and project these points to each input frame's pixel space. The feature encoding for each point is then obtained by bilinearly interpolating the 2D DINO feature map. This yields a sequence of 3D DINO feature volumes, which are subsequently aggregated across time for a single 3D feature volume $\mathcal{F}_{\text{3D}} \in \mathbb{R}^{Z \times Y \times X \times 384}$, where $Z = X = 96, Y = 32$ are the spatial resolutions of the voxel in our Y-up canonical frame. We simply use average pooling across time as the temporal aggregation method.

BEV projection. Directly reasoning in the 3D feature space is expensive and in many cases sub-optimal (as ablated later). We hence project the 3D feature volume obtained above to the "Bird's Eye View" (BEV) by "squeezing" the up-axis (Y axis), following [12, 13]. The squeezing is done with an MLP that projects the flattened up-channel dimension ($384 \times Y$) to the same-sized channel dimension (384). After this step, we end up with a BEV feature map $\mathcal{F}_{\text{BEV}} \in \mathbb{R}^{Z \times X \times 384}$.

BEV Net. $\mathcal{F}_{\text{BEV}}$ is the condensed feature embedding on which we perform the bulk of the computation. Following previous designs [12, 13, 43], our BEV Net consists of 11 sequential `BEV_modules`, where each module applies a 2D convolution, a LayerNorm, and an MLP with a GELU activation. The hidden dimension starts at 384 and doubles twice in the middle, reaching the final feature dimension of 1540 while reducing the spatial dimension to 3 × 3.

Trajectory prediction. We first perform a spatial average pooling on the features from the BEV Net, and then use a 3-layer MLP with LayerNorm and GELU activation to get the predicted future head pose sequence $\hat{\mathcal{H}}_{T_1+1:T_1+T_2}$.

Loss function. We supervise the model with combined L1 losses on the translations and rotations, following [26]:

$$\mathcal{L} = \frac{1}{T_2} \cdot \sum_{t=T_1+1}^{T_1+T_2} \lambda_{\text{trans}} \cdot ||\mathbf{t}_t - \hat{\mathbf{t}}_t||_1 + \lambda_{\text{rot}} \cdot ||\mathbf{R}_t \hat{\mathbf{R}}_t - \mathbf{I}||_1 \quad (1)$$

where R is the rotation matrix converted from the 6D rotation representation, and I is the identity matrix. In our experiments we use $\lambda_{\text{trans}} = \lambda_{\text{rot}} = 1$.

3.3. Implementation Details

Head-centered canonical frame. Because we do not input the past head poses to the model (they are only used in the unprojection process), we need to define a canonical frame relative to the current head pose $\mathbf{h}_{T_1}$. Such canonical frames have been widely adopted in prior works [11, 26, 33, 49, 65]. Following [33], we also define our frame to be parallel to the ground plane and facing forward, but centered on the head instead. This lets the model operate in a space facing the current heading direction and predict future head poses in a relative sense.

Training details. We use the AdamW [28] optimizer, and apply a weight decay of 0.05 to all model parameters except biases. We train our model for 700k steps, and use the OneCycle learning rate scheduler [55] with the linear annealing strategy and `pct_start` set to 0.05. We use a batch size of 4. The training concludes in about 4 days on a single NVIDIA RTX A6000 GPU.

4. Aria Navigation Dataset (AND)

As discussed in Sec. 2, there are no suitable datasets for our task to the best of our knowledge. We hence design a data collection pipeline to collect our own dataset, which we name as the Aria Navigation Dataset (AND). We next present key data collection steps and dataset statistics.

4.1. Data Collection Pipeline

Hardware. Our data collection hardware consists of only a pair of the Project Aria glasses [7], which has several key advantages of being lightweight, non-intrusive, cheap, and easy-to-set-up compared to prior works that deploy self-built sensor suites [36, 63] or teleoperate robots [17].

Recording process. Project Aria comes with a mobile app named Aria Studio that allows easy data recording by interacting with the mobile app once prior to each recording session. The app provides selections on recorded data modalities, and in our pipeline, we activate the RGB, SLAM (two monochrome cameras), and eye-tracking cameras, as well as IMU sensor, barometer, and GPS. The cameras all operate at 20fps. Before each recording session, the human subject selects this saved recording profile and starts recording while walking around, without pre-defined instructions or scripts. To capture consistent information-gathering behavior, we instruct the subjects to follow a careful navigation behavior, *e.g.* always checking for passing vehicles before crossing roads.

Data processing. The raw recorded data comes in a compressed format called VRS [†]. We run the Aria Machine Perception Services [‡] to get processed data modalities including 6D head pose trajectories and scene point clouds. The raw RGB frames are distorted due to the fisheye camera, so we undistort them to use as input to our model. The original sequences are further segmented into (T_1+T_2)-long clips in a sliding window fashion, with a stride and dilation factor of 6 frames. At 20fps, each clip covers $(8+8-1)\times 6/20 = 4.5$ seconds. During SLAM, points on the dynamic objects are filtered. We apply another filtering process on the reconstructed point cloud to remove noisy points.

Privacy. We have taken measures to follow Project Aria research guidelines. We also use the SOTA de-identification algorithm [48] to blur faces in all videos.

4.2. Dataset Statistics

Locations. Since we want to capture real-world navigation scenarios where humans need to avoid collision with both static and dynamic obstacles, we selected diverse locations with dense traffic both indoors and outdoors. We picked 18 densely populated places from university campuses, city downtowns, parks, and so on. Many of these locations are expansive, providing great diversity on captured data. We specifically chose times when dense traffic usually happens for data recording, *e.g.* after classes. We also diversified the time-of-the-day distribution.

Data scale. We recorded about 4 hours of data, resulting in 274k RGB frames and 36k clips after processing.

5. Experimental Results

We compare LookOut to baselines quantitatively in Sec. 5.1.1. In Sec. 5.1.2, we ablate our key design choices. We then present qualitative evaluation results in Sec. 5.2, which showcase the diverse behaviors our model learns in real-world navigation scenarios. Finally, we investigate failure cases from our model and discuss limitations in Sec. 5.3. All results are obtained on a held-out set with environments unseen during training. We encourage readers to check our project webpage, which contains video versions of Fig. 3 obtained by continuously rolling out our model given each incoming frame, similar to how it would operate in practice.

5.1. Quantitative Evaluation

Metrics. We first evaluate head pose prediction accuracy through the same error function we used for training, *i.e.* the L1 losses on translation (`L1_trans`) and rotation (`L1_rot`). In order to measure collision with the environment, we define a non-collision score for the static (`Col_stt_k`) and dynamic (`Col_dyn_k`) obstacles, respectively. The score measures the percentage of predictions that are at least k centimeters away from the closest obstacle. For static obstacles, we measure the closest distance from the predicted head translation to the SLAM reconstructed point cloud. While for dynamic obstacles, we first use a monocular metric depth estimation method Depth Pro [1] to estimate a depth map for each frame in our dataset, and subsequently use DINOv2 + Mask2Former segmentation head [38] to get per-frame semantic segmentation masks. We then take the minimum estimated metric depth values among all pixels labeled "person" as the closest distance. `Col_*_avg` is the average value over all $k \in \{15, 25, 35\}$. Note that the non-collision score is a rough proxy for collision avoidance, and we also report the values for ground-truth sequences (`GT`) for reference.

5.1.1. Comparison with Baselines

Baselines. Since we study a novel problem, there is no directly comparable prior work. As mentioned, the closest prior works to ours are EgoNav [63] and EgoCast [8]. EgoNav is an Arxiv preprint that did not release code. We hence adapt EgoCast to our setting. The core part of EgoCast is a transformer-based forecasting module that predicts future 3D full-body poses given past full-body poses and optionally the past egocentric video. To deal with the issue that full-body poses are often not available in practice, it further implements an estimation module to estimate the current frame's full-body pose from past 6D head poses and egocentric video. Both stages are supervised with 3D full-body poses, which our dataset does not have. We hence repurpose the forecasting module to take past head poses (instead of full-body poses) and the egocentric video, and predict future

[†] https://facebookresearch.github.io/vrs/docs/Overview/

[‡] https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps

Method	L_1_trans ↓	L_1_rot ↓	Col_stt_15 ↑	Col_dyn_15 ↑	Col_stt_25 ↑	Col_dyn_25 ↑	Col_stt_35 ↑	Col_dyn_35 ↑
Const_Vel	0.41	0.77	85.5	91.2	80.0	81.3	74.1	73.1
Lin_Ext	0.45	1.21	86.5	92.3	77.6	82.1	73.3	72.9
EgoCast [8]	0.34	0.63	90.5	94.6	84.6	86.2	77.4	77.8
Ours	**0.17**	**0.16**	**91.3**	**97.2**	**85.6**	**90.3**	**79.9**	**83.1**
GT	0	0	92.7	97.7	88.9	93.0	83.6	85.1
A^*+Lin_Ext	0.24	1.21	**98.8**	82.4	**100.0**	76.5	**100.0**	61.9
Ours (+goal)	**0.11**	**0.15**	91.7	**97.2**	86.3	**91.4**	82.0	**84.6**

Table 1. **Comparison with baselines.** Our model outperforms comparable methods on both trajectory prediction and collision avoidance.

head poses too. We then remove the estimation module. We train it on the same training split as our model.

We additionally implement the following baselines that operate on past head poses: (1) Constant Velocity (`Const_Vel`) that uses the linear and angular velocity calculated from the last two input steps to extrapolate future head translations and rotations, (2) Linear Extrapolation (`Lin_Ext`) that fits a linear regression model for the past translation and rotation sequences and predicts into the future, and (3) A^*+Linear Extrapolation (`A*+Lin_Ext`) that uses linear extrapolation for rotations, but implements an A^* algorithm for translations. Specifically, we discretize the space by turning the SLAM reconstructed point cloud into an occupancy grid, using the same spatial resolution as our model. We then take the ground-truth head translation from the last step in the future $T_1 + T_2$ as the goal. We also define a maximum velocity that roughly matches the human movement capacity. We use a variant of our model that also takes such goal position as input (directly concatenated to the final MLP) for a fair comparison with this baseline.

Analyses. The comparison results are reported in Tab. 1. In the no-goal setting (top), our model achieves the best performance across all metrics, predicting accurate head poses while avoiding collision with both static and dynamic obstacles reliably. When provided with the goal, `A*+Lin_Ext` achieves near-perfect non-collision scores for static obstacles, because they are explicitly modeled with the scene occupancy (where each voxel grid represents about 600cm^3 of space) and the search algorithm basically guarantees a path around occupied regions. However, this baseline does poorly in avoiding dynamic obstacles since they are not represented in the point cloud.

5.1.2. Ablation Study

We ablate input data modalities and key model designs and summarize the results in Tab. 2.

Method	L_1_trans ↓	L_1_rot ↓	Col_stt_avg ↑	Col_dyn_avg ↑
PCD only	0.40	0.88	83.2	84.6
RGB+PCD	0.17	0.14	**87.8**	90.1
Depth only	0.22	0.23	87.0	**91.6**
RGB+Depth	**0.15**	**0.13**	87.4	91.4
w/o DINO	0.35	0.67	84.5	85.3
2D Only	0.26	0.44	84.9	86.2
3D Conv	**0.17**	0.19	**85.6**	89.9
Ours	**0.17**	**0.16**	**85.6**	**90.2**
GT	0	0	88.4	91.9

Table 2. **Ablation study.** Each design choice helps performance.

Multi-modal support. Our model can be easily extended to incorporate additional sensor modalities, *e.g.* depth and point cloud. Specifically, we first convert depth to a point cloud, and then turn it into an occupancy voxel and concatenate with the 3D DINO volume $\mathcal{F}_{3\text{D}}$. As expected, incorporating these modalities directly relevant to obstacle proximity improves non-collision metrics, which aligns with the findings in previous works [40]. Using depth especially helps with avoiding dynamic obstacles because the SLAM reconstructed point cloud only contains static objects.

Model design. We first validate the effectiveness of the DINO feature encoding by ablating a variant that unprojects raw RGB frames without passing through DINO (`w/o DINO`). As shown, DINO features contribute significantly to our model's performance due to its strong semantic feature encoding capabilities. We next inspect the impact of having the intermediate 3D feature space, for which we ablate a variant that temporally pools over 2D DINO features $\mathcal{F}_{2\text{D}}$ instead (`2D Only`). We can see that the 3D feature space improves performance by granting an explicit geometric notion to the features. Finally, we investigate whether the BEV projection is beneficial by comparing against directly applying 3D convolutions on $\mathcal{F}_{3\text{D}}$ (`3D Conv`). This variant performs on par to our model, but is more computationally expensive due to 3D convolutions.

5.2. Qualitative Evaluation

Diverse model behaviors. We evaluate if our trained model demonstrates desirable behaviors through visual inspection. Specifically, we are interested in seeing (1) if our model predicts trajectories that are free from collision with static and dynamic obstacles, (2) if our model learns human-like information-gathering behaviors, and (3) where our model fails. For this purpose, we visualize our model's

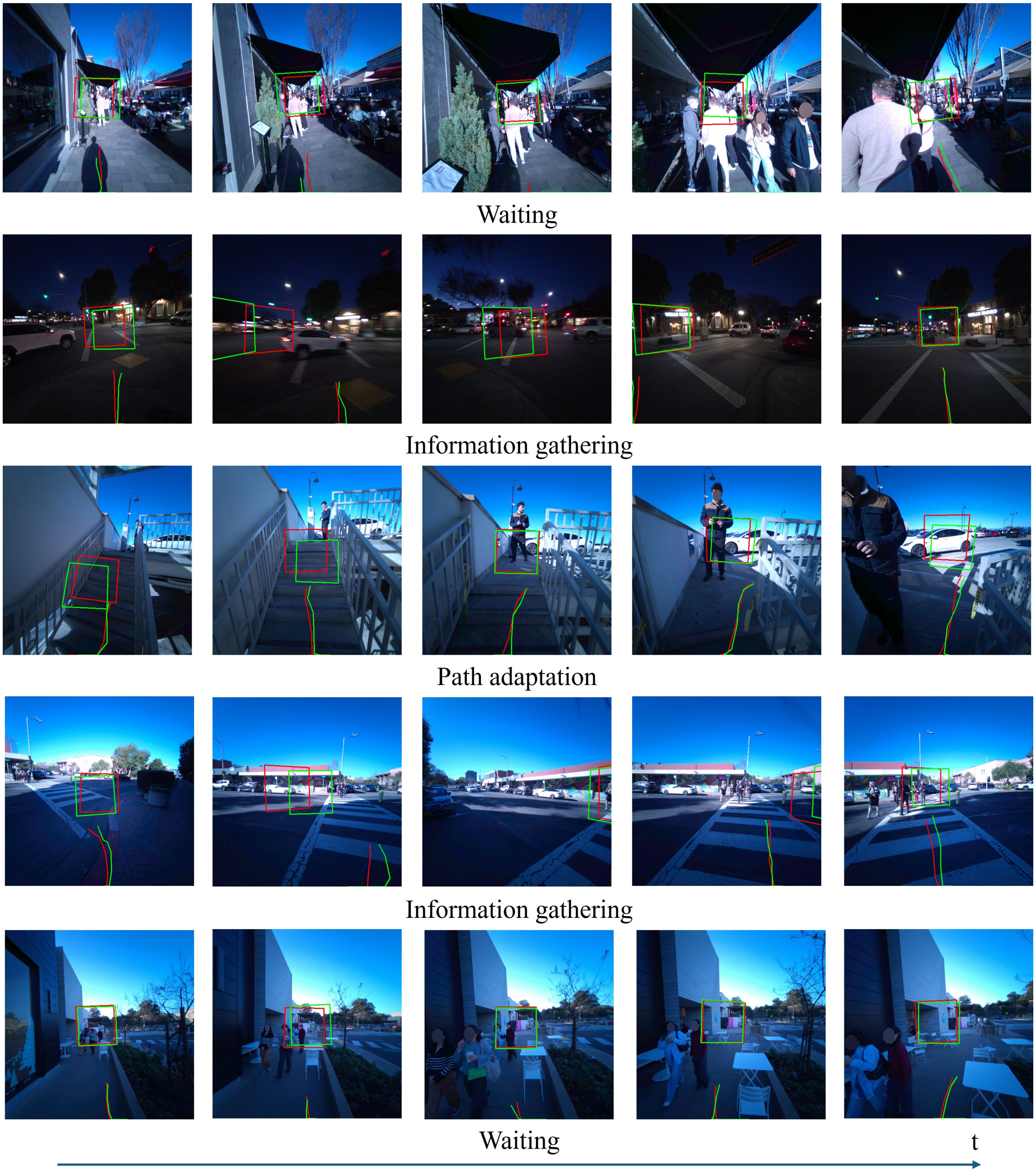

Figure 3. **Visualizations of model behaviors**. We provide five examples from the held-out set with model predictions (red) and ground-truths (green). For each example, we show five frames and a text describing the model behavior below. The translation visualizations (curves) are obtained by projecting the values to the ground and then to the image plane. The rotations (squares) are visualized by projecting the viewing frustums to the image plane. We show the full future sequences for translations while only the next step for rotations for clarity.

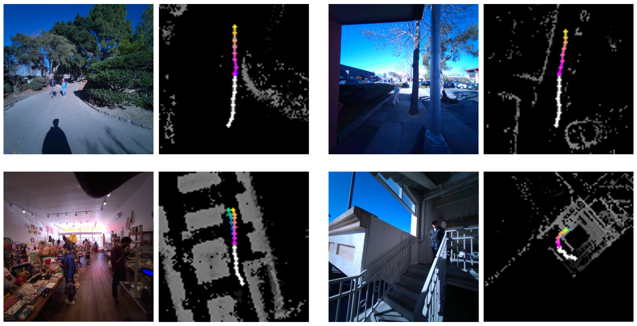

Figure 4. **BEV visualizations**. We show four examples, each with a sampled RGB frame on the left and the BEV visualization of trajectories on the right. The white curve denotes the past, blue-green denotes the ground-truth future, and pink-orange represents the predicted future. Color coding depicts the order of time progression. The trajectories are overlaid on a BEV representation of the scene point cloud for visualization. Note that only the translation components of the trajectories are shown here.

predictions and ground-truths in 2D, and overlay them over the image observations. We show such visualizations for a few samples in Fig. 3 and more on the project webpage. It can be seen that our model forecasts collision-free paths both around static and dynamic obstacles. Our model also learns the information-gathering behavior that humans demonstrate in the training data, such that it predicts head rotations that check potentially useful information (*e.g*. road conditions) for navigation. We also observe other interesting behaviors from the model. In the first and fifth examples, the model learns to wait when there is no easily traversable path available. In the third example, the model adapts its predictions based on new visual cues it observes (the predicted path shifts from the center to the right after the person appears).

BEV visualizations. We additionally show translation visualizations from the BEV in Fig. 4. We overlay the trajectories on top of a BEV representation of the static environment, which is obtained by converting the scene point cloud to an occupancy grid and then to a height map. The height map stores for each pixel, the maximum height of all occupied grids along the up-axis. As seen again from these visualizations, the predicted trajectories from our model satisfy environmental constraints in diverse scenarios.

5.3. Failure Cases and Limitations

We identify failure cases of our model and provide the visualizations in Fig. 5. A major limitation of our model is its lack of generative modeling capabilities and hence may struggle when a multi-modal future is possible. In the first example of Fig. 5, it is possible to go either left or right to avoid collisions with the incoming pedestrians, in which case our model shall regress to the mean of these multiple possibilities. Only when the human subject in this case clearly walks to their right side, our model is able to regress to a plausible future. A next step is thus to leverage generative models to learn such multimodal distributions, such as diffusion models [15, 57]. In the second case, the human subject looks down to check the position of the rail in the middle time step to avoid tripping. However, our model does not make such predictions because rails have never appeared in our training set. Expanding our training set to include more diverse scenarios would be a promising solution to this problem.

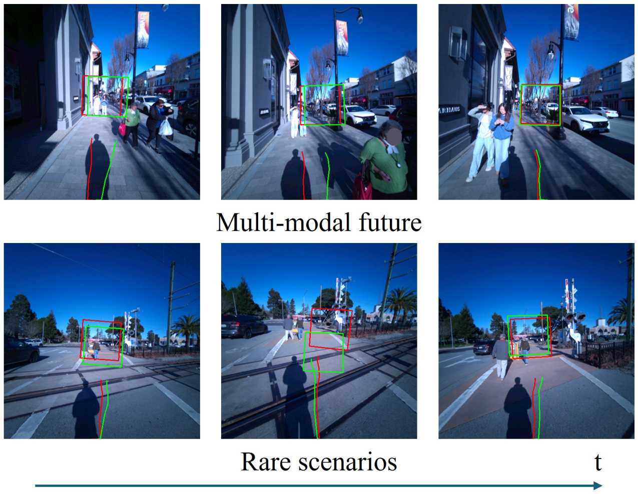

Figure 5. **Failure cases**. We show two examples from the held-out set with a description of the failure below each.

6. Conclusion

In this paper, we make steps towards a real-world deployable humanoid navigation policy by making a number of contributions. First, we introduce a novel task of predicting the future 6D head pose trajectory from the past posed egocentric video, under the presence of both static and dynamic obstacles. This task formulation allows the model to learn to not only plan collision-free paths but also learn human-like information-gathering behaviors. Second, we propose a model that leverages pre-trained DINO feature encoders and a parameter-free unprojection strategy to effectively solve this task. Next, we design a data collection pipeline that uses only a pair of Project Aria glasses as the data capture device. This pipeline is easily scalable and allows us to collect a 4-hour real-world navigation dataset with ease. Our dataset spans 18 places with diverse and dense traffic, providing the community with a valuable resource. Through extensive experiments, we demonstrate that our model learns diverse behaviors that are useful for real-world navigation tasks, and surpasses baselines across all metrics. Finally, we discuss the failure cases and limitations of our model as well as directions for future work.

Acknowledgments

This work was supported by ARL under grant number W911NF-21-2-0104, a Vannevar Bush Faculty Fellowship, HAI, and Meta Reality Labs.

References

[1] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 5

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 3

[3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 2

[4] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024. 1

[5] Ricson Cheng, Ziyan Wang, and Katerina Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 3

[7] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 3, 5

[8] Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbelaez. Egocast: Forecasting egocentric human pose in the wild. *arXiv preprint arXiv:2412.02903*, 2024. 1, 3, 5, 6

[9] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023. 2

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 3

[11] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. *arXiv preprint arXiv:2409.13426*, 2024. 3, 4

[12] Adam W Harley, Shrinidhi K Lakshmikanth, Fangyu Li, Xian Zhou, Hsiao-Yu Fish Tung, and Katerina Fragkiadaki. Learning from unlabelled videos using contrastive predictive neural 3d mapping. *arXiv preprint arXiv:1906.03764*, 2019. 3, 4

[13] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023. 3, 4

[14] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023. 1

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 8

[16] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pages 277–294. Springer, 2024. 3

[17] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7 (4):11807–11814, 2022. 1, 2, 3, 5

[18] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024. 1, 2

[19] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022. 3

[20] Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, and Devendra Singh Chaplot. Navigating to objects specified by images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10916–10925, 2023. 2

[21] Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, and Devendra Singh Chaplot. Navigating to objects specified by images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10916–10925, 2023. 1, 2

[22] Iori Kumagai, Mitsuharu Morisawa, Shin'ichiro Nakaoka, and Fumio Kanehiro. Efficient locomotion planning for a humanoid robot with whole-body collision avoidance guided

by footsteps and centroidal sway motion. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 251–256. IEEE, 2018. 3

[23] Chung-Hsien Kuo, Hung-Chyun Chou, Shou-Wei Chi, and Yu-De Lien. Vision-based obstacle avoidance navigation with autonomous humanoid robots for structured competition problems. *International Journal of Humanoid Robotics*, 10(03):1350021, 2013. 3

[24] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. Egogen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14497–14509, 2024. 3

[25] Heng Li, Minghan Li, Zhi-Qi Cheng, Yifei Dong, Yuxuan Zhou, Jun-Yan He, Qi Dai, Teruko Mitamura, and Alexander Hauptmann. Human-aware vision-and-language navigation: bridging simulation to reality with dynamic human interactions. *Advances in Neural Information Processing Systems*, 37:119411–119442, 2025. 1, 2

[26] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 3, 4

[27] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[29] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 3

[30] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2024. 3

[31] Daniel Maier, Maren Bennewitz, and Cyrill Stachniss. Self-supervised obstacle detection for humanoid navigation using monocular vision and sparse laser data. In *2011 IEEE international conference on robotics and automation*, pages 1263–1269. IEEE, 2011. 3

[32] Aashi Manglik, Xinshuo Weng, Eshed Ohn-Bar, and Kris M Kitanil. Forecasting time-to-collision from monocular video: Feasibility, dataset, and challenges. In *2019 ieee/rsj international conference on intelligent robots and systems (iros)*, pages 8081–8088. IEEE, 2019. 3

[33] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, pages 903–913. IEEE, 2024. 4

[34] Siddarth Narasimhan, Aaron Hao Tan, Daniel Choi, and Goldie Nejat. Olivia-nav: An online lifelong vision language approach for mobile robot social navigation. *arXiv preprint arXiv:2409.13675*, 2024. 3

[35] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9890–9900, 2020. 3

[36] Duc M Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7442–7447. IEEE, 2023. 1, 3, 5

[37] Koichi Nishiwaki, Joel Chestnutt, and Satoshi Kagami. Autonomous navigation of a humanoid robot over unknown rough terrain using a laser range sensor. *The International Journal of Robotics Research*, 31(11):1251–1262, 2012. 3

[38] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1, 3, 5

[39] Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation. *arXiv preprint arXiv:2310.07889*, 2023. 1, 2

[40] Boxiao Pan, Bokui Shen, Davis Rempe, Despoina Paschalidou, Kaichun Mo, Yanchao Yang, and Leonidas J Guibas. Copilot: Human-environment collision prediction and localization from egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5262–5272, 2023. 3, 6

[41] Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dinesh Manocha, and Xuesu Xiao. Social-llava: Enhancing robot navigation through human-language reasoning in social spaces. *arXiv preprint arXiv:2501.09024*, 2024. 1, 3

[42] Chengyang Peng, Victor Paredes, Guillermo A Castillo, and Ayonga Hereid. Real-time safe bipedal robot navigation using linear discrete control barrier functions. *arXiv preprint arXiv:2411.03619*, 2024. 1, 2

[43] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020. 4

[44] Davide Plozza, Steven Marty, Cyril Scherrer, Simon Schwartz, Stefan Zihlmann, and Michele Magno. Autonomous navigation in dynamic human environments with an embedded 2d lidar-based person tracker. In *2024 IEEE Sensors Applications Symposium (SAS)*, pages 1–6. IEEE, 2024. 3

[45] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9982–9991, 2020. 2

[46] Ilija Radosavovic, Sarthak Kamat, Trevor Darrell, and Jitendra Malik. Learning humanoid locomotion over challenging terrain. *arXiv preprint arXiv:2410.03654*, 2024. 1

[47] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[48] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, et al. Egoblur: Responsible innovation in aria. *arXiv preprint arXiv:2308.13093*, 2023. 5

[49] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 4

[50] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 3

[51] Kohtaro Sabe, Masaki Fukuchi, J-S Gutmann, Takeshi Ohashi, Kenta Kawamoto, and Takayuki Yoshigahara. Obstacle avoidance and path planning for humanoid robots using stereo vision. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, pages 592–597. IEEE, 2004. 3

[52] José Ricardo Sánchez-Ibáñez, Carlos J Pérez-del Pulgar, and Alfonso García-Cerezo. Path planning for autonomous mobile robots: A review. *Sensors*, 21(23):7898, 2021. 3

[53] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018. 3

[54] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 3

[55] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 4

[56] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 2024. 1, 3

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 8

[58] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3

[59] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019. 3

[60] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453. IEEE, 2022. 3

[61] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*, pages 367–385. Springer, 2024. 3

[62] Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2595–2603, 2019. 3

[63] Weizhuo Wang, C Karen Liu, and Monroe Kennedy III. Egonav: Egocentric scene-aware human trajectory prediction. *arXiv preprint arXiv:2403.19026*, 2024. 1, 3, 5

[64] Xiaomeng Xu, Yanchao Yang, Kaichun Mo, Boxiao Pan, Li Yi, and Leonidas Guibas. Jacobinerf: Nerf shaping with mutual information gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16498–16507, 2023. 3

[65] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 3, 4

[66] Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5543–5550. IEEE, 2024. 1, 2

[67] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 3

[68] Xiaoming Zhao, Harsh Agrawal, Dhruv Batra, and Alexander G Schwing. The surprising effectiveness of visual odometry techniques for embodied pointgoal navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16127–16136, 2021. 2

[69] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 3