注意：真实世界的人形自我中心导航

Boxiao Pan　　Adam W. Harley　　C. Karen Liu*　　Leonidas J. Guibas*

Stanford University

Abstract

从自我中心的观察中预测无碰撞的未来轨迹的能力在仿人机器人、虚拟现实/增强现实和辅助导航等应用中至关重要。在这项工作中，我们介绍了从自我中心视频预测未来 6D 头部姿态序列这一具有挑战性的问题。特别是，我们预测头部平移和旋转以学习通过头部转动事件表达的主动信息收集行为。为了解决这个任务，我们提出了一种基于时间聚合的 3D 潜在特征推理框架，该框架对环境的静态和动态部分进行几何和语义约束建模。受此领域训练数据缺乏的启发，我们进一步贡献了一个使用 Project Aria 眼镜的数据采集管道，并通过这种方法展示了一个收集到的数据集。我们的数据集被称为 Aria 导航数据集（AND），包含 4 小时用户在现实世界场景中导航的记录。它包括各种各样的情况和导航行为，为学习现实世界的自我中心导航策略提供了一个宝贵的资源。广泛的实验表明，我们的模型学会了像人类一样的导航行为，如等待/减速、重新规划路线以及环顾四周观察交通，并能够推广到未见过的环境。访问我们的项目网页 https://sites.google.com/stanford.edu/lookout。

1. 介绍

在现实世界中从以自我为中心的观察中安全导航是人类具备的能力，但机器学习起来却极其困难。这主要是由于实际场景中存在的多样且复杂的状况。这样的能力对于包括类人机器人 [42]、虚拟现实/增强现实 [8] 和辅助导航 [63] 等各种应用都至关重要。

许多研究工作从不同的角度解决了这个问题。视觉语言导航（VLN）[18, 21, 25, 39, 66] 关注目标定位和规划长期目标导向路径，通常在模拟静态环境中进行。机器人社交导航 [14, 17, 36, 41, 56] 学习在动态环境中的社会合规导航策略。这些工作一般针对轮式或腿式导航机器人，其动作和观察分布与人类体型有极大差异。最近，一些研究调查了预测人类轨迹 [63] 或全身姿态 [8] 的以自我为中心的导航。然而它们假设环境是静态的。

尽管取得了这些进步，可部署到现实世界的人形自我中心导航策略仍然具有挑战性。首先，缺乏在动态环境中进行人形导航的方法。其次，现有方法忽略了类人类导航的一个重要方面，即通过转头主动收集信息。人们经常转动头部寻找有用的信息。例如，在过马路前我们会向旁边看以检查是否有过往车辆，在上下路缘时会向下看等等。这种能力有助于现实世界的部署，部分原因是摄像机的有限视场。最后，由于将人形机器人部署到现实世界中的困难，我们没有一种方法可以大规模收集多模态标注训练数据。

在这项工作中，我们从这三个方面朝着一个现实世界可部署的类人导航策略迈进。为了解决三维动态场景感知的挑战，我们提出了一种将每帧的 DINO [2, 38] 特征反投影到 3D，并在时间上聚合 3D 特征体积以获得对环境所施加几何约束的整体理解的模型。此外，通过在包含大量动态障碍物（i.e. 行人和车辆）的数据集上进行训练，我们的模型有效地学习了绕过静态和动态物体导航的能力。为了建模主动信息收集，我们设计框架预测 3D 头部旋转外加平移（i.e. 6D 头部姿态预测），这可以用来计算通常输入给人形机器人的速度命令 [4, 46, 47]。此外，在我们的数据收集过程中，我们要求人类参与者遵循一个谨慎的信息收集策略，e.g. 始终在过马路前寻找过往车辆。为了解决大规模收集有用

* 平等指导

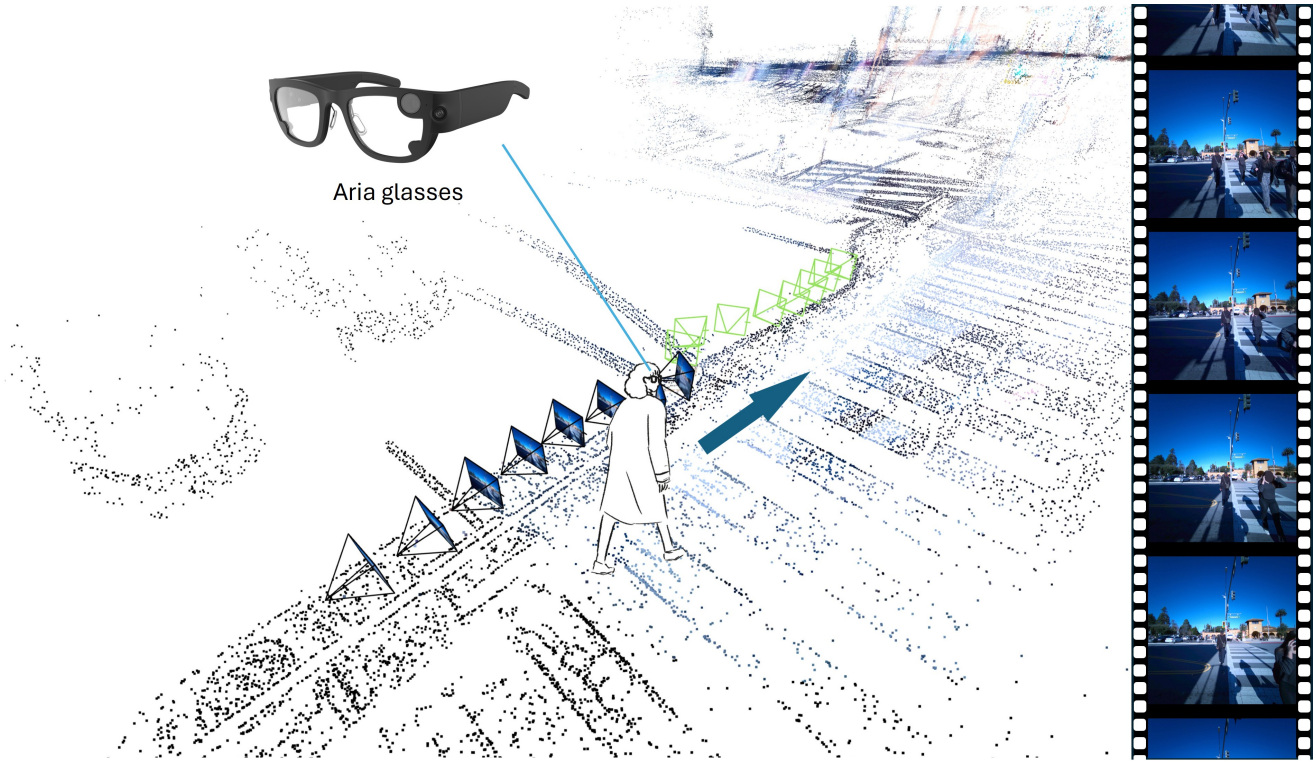

图 1. **问题表述**. 给定一个以自我为中心的视频（用黑色轮廓表示的棱台，右侧详细展示了帧），我们的模型预测未来的一系列 6D 头部姿态（用绿色轮廓表示的棱台）。我们设计了一个使用 Project Aria 眼镜的数据收集流程，并通过这种方式收集的数据集对模型进行训练。这个问题包含了现实世界的导航挑战，包括与静态和动态障碍物的避碰，以及类似人类的信息采集行为（e.g. 例如过马路时向两侧看）。点云用于可视化但不是模型的输入。

数据的挑战，我们提出了一种使用一对 Project Aria 眼镜 [7] 作为数据收集工具的数据收集管道。这个管道使自然的人类导航演示收集成为可能而不引起注意 [42]。与需要仔细安装各种传感器 [42] 或遥操作机器人 [17] 的传统收集管道不同，我们的管道非常容易设置，在每个录制会话开始时只需几秒钟即可完成设置，同时提供包括 RGB 视频、音频、眼动追踪、SLAM 重建头部姿势和点云在内的各种数据模式。因此，它以最小的努力提供了扩展数据收集过程的方法。使用这个管道，我们收集了一个包含 4 小时现实世界导航会话的数据集，涵盖了 18 个密集人口地区。

总结，我们做出了以下贡献：(1) 我们在存在静态和动态障碍物的情况下，引入了从姿态自我中心观察预测 6D 头部姿态轨迹这一具有挑战性的任务。(2) 我们提出了一种名为 LookOut 的模型，该模型在时间上聚合未投影的 3D DINO 特征以实现语义和几何理解，在解决此任务中证明了其有效性。(3) 我们贡献了一个数据收集管道，利用一对 Project Aria 眼镜，并且需要极小的努力，提供了一种轻松扩大数据收集规模的方法。(4) 我们使用该管道收集了一个包含 4 小时真实世界导航会话的数据集，覆盖了 18 个具有密集和多样化交通的地点。

2. 相关工作

视觉语言导航． 视觉语言导航（VLN）任务的研究可以说是围绕具身导航的计算机视觉领域中最普遍的。该任务大致定义为导航到指定的目标位置。根据目标的具体描述，此任务有几个变体：Point-Goal Navigation (PointNav) [68], Object-Goal Navigation (ObjectNav) [9, 21, 66], Image-Goal Navigation (ImageNav) [20], Language-Goal Navigation (LangNav) [39, 45] 以及 Audio-Visual Navigation [3]。一些工作也提出了统一多个规范的框架 [18, 25] 。这些工作通常是在模拟环境中开发的，专注于长期路径规划，在这种情况下不存在或只有简单的动态障碍物 [25] 。在这项工作中，我们关注的是短期导航，其主要目标是无

碰撞移动。

机器人社会导航 。经典方法广泛研究了给定几乎完美环境知识的机器人路径规划问题 [52]，提出了数学解决方案。与我们的工作更相关的是，有关于机器人自我中心导航的研究。机器人社交导航旨在预测一个无碰撞路径 [34, 44, 56] 或更高层次指令 [32, 41] 给定诸如 RGB、LiDAR 和里程计这样的自我中心感知输入。此类任务的常见数据集是通过遥控操作机器人 [17] 或人类收集者佩戴传感器套件来收集的 [36]。这些研究通常针对轮式或腿足导航机器人。这类机器人的行动和观察与类人机器人截然不同，原因是它们体积较小、速度更快且形态各异。另一方面，对于人形机器人 [22, 23, 31, 37, 51] 的研究通常设计使用激光和立体视觉输入的经典方法来应对简化的环境。

以自我为中心的人类导航 . 在以自我为中心的人体运动估计 [11, 16, 26, 35, 50, 59, 65, 67] 方面取得了显著进展，而预测考虑环境约束的未来运动或轨迹的研究则相对较少。COPILOT[40] 从多视角以自我为中心的视频中预测人与环境的碰撞，并以碰撞标签和热图的形式呈现。EgoNav[63] 使用扩散模型，根据胸部安装的 RGBD 摄像头输入和过去的轨迹预测未来的轨迹。EgoCast[8] 根据头部安装的 RGB 摄像头输入和过去的头部轨迹预测未来的全身姿态。值得注意的是，EgoNav 主要关注导航，而 EgoCast 研究多样化的社会和专业人类活动。然而，这三项研究都假设环境是静态的，并且没有学习在现实世界场景中至关重要的主动信息收集行为。此外，我们贡献了一个可以轻松大规模部署的数据管道。

以自我为中心的人类数据集。为了研究这个任务，我们需要一个记录现实世界人类导航场景的数据集，在这些场景中同时存在静态和动态障碍物，并且提供以自我为中心的 RGB 视频、6D 头部姿态以及优选地还提供用于碰撞检测的场景点云数据模式。传统的自我为中心的视频数据集 [6, 10, 27, 53] 通常是通过单目设置捕捉的，因此没有相机姿态标注。这些数据集中的活动也多种多样，并不专注于导航。一些研究 [24, 40] 提出了合成数据生成管道来模拟虚拟人类在合成场景中行走。然而，由于人体运动生成方法的限制，产生的动作简单且不自然。自动驾驶数据集如 Waymo Open [58] 全面涵盖了现实世界的交通场景，并为行人和车辆提供了密集的 3D 跟踪，但它们没有记录行人的自我为中心的数据。最近，Project Aria 眼镜 [7] 作为记录自我为中心数据的一种方便且自然的方式出现。特别是，Aria 机器感知服务（MPS）提供了一种简单且高度优化的方法来获取准确的 6D 相机（头部）姿态轨迹、环境点云、眼睛注视等。Meta 发布了几个由 Project Aria 收集的数据集，其中 Aria Everyday Activities (AEA) [29] 和 Nymeria [30] 记录了多样的室内和室外活动。然而，它们要么仅包含单人活动 [29]，要么只有在协作活动中才有多名参与者 [30]，因此没有捕捉到潜在碰撞代理的真实世界导航场景。

3. 方法

3.1. 问题表述

我们通过 Fig. 1 来说明我们的问题公式。给定一个以自我为中心的视频 $\mathcal{X} \in \mathbb{R}^{T_1 \times H \times W \times 3}$ 和 $\mathcal{H}_{1:T_1} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{T_1}\} \in \mathbb{R}^{T_1 \times 9}$，我们的目标是预测未来短时间内 6D 头部姿态序列，i.e. $\mathcal{H}_{T_1+1:T_1+T_2} = \{\hat{\boldsymbol{h}}_{T_1+1}, \hat{\boldsymbol{h}}_{T_1+2}, \cdots, \hat{\boldsymbol{h}}_{T_1+T_2}\} \in \mathbb{R}^{T_2 \times 9}$。头部姿态，也就是相机的姿态，被参数化为 $\boldsymbol{h}_t = [\boldsymbol{t}_t | \boldsymbol{r}_t]$，其中旋转部分 $\boldsymbol{r}_t$ 采用 6D 连续旋转表示法 [69]。在整个实验过程中，$T_1 = T_2 = 8$。头部姿态在以头部为中心的规范框架中定义，该框架指定于 Sec. 3.3。

3.2. 模型

我们模型需要具备的核心功能是从单一单目视频流中提取周围环境的语义和几何信息。这促使我们考虑以下问题：(1) 用于语义建模的强大视觉编码器，以及 (2) 一种在 3D 空间中聚合信息并推理的方法。这些问题激发了我们的模型的关键设计选择。对于 (1)，我们使用预训练的 DINO [2, 38] 编码器来提取每帧特征图，因为其具有强大的开放式词汇语义编码能力 [19, 60, 61, 64]。对于 (2)，我们采用在多个对象和场景表示模型中使用的一种名为“无参数反投影” [5, 12, 13, 54, 62] 的策略。具体来说，我们对定义在规范框架中的每个 3D 坐标进行双线性采样以获得子像素的 2D DINO 特征，从而得到一个 3D DINO 特征体积。然后我们将所有时间步骤的体积进行时间上的聚合。这些设计选择赋予了我们的模型强大的语义和几何推理能力，而无需依赖诸如深度和激光雷达等显式的几何感知输入。我们在 Fig. 2 中展示了该模型，并将在下面详细描述每个组件。

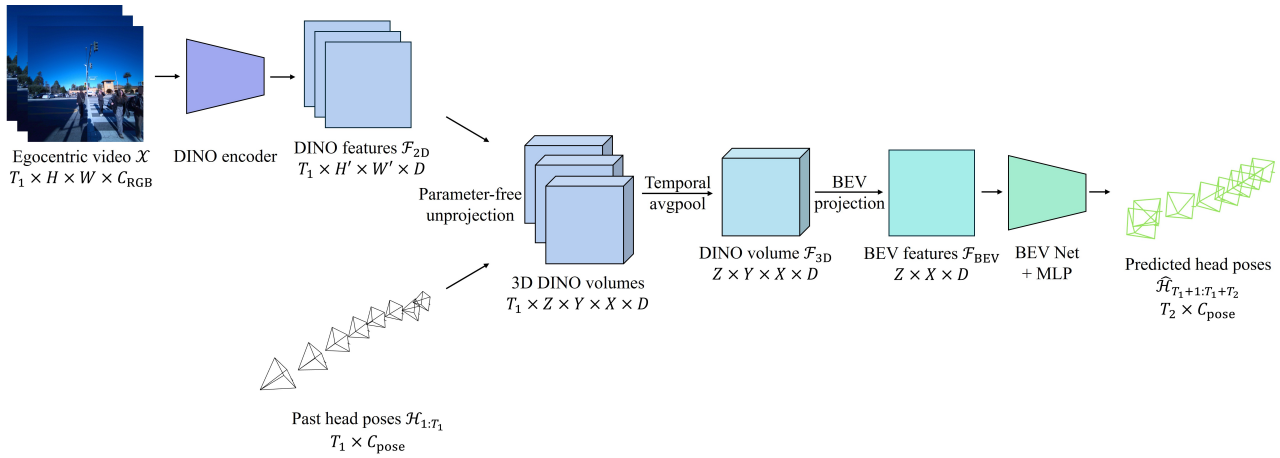

图 2. **注意 架构**. 给定一个带有姿态的以自我为中心的视频，我们使用预训练编码器获取逐帧的 DINO 特征，并将其反投影到 3D 空间进行时间聚合。然后将聚合后的特征投影到 BEV 进行进一步处理，并最终用于预测未来的头部姿态。

DINO **特征编码**. 我们使用预训练的 DINO2 [38] 变体 dinov2_vits14_reg *，并将其应用于每个输入帧（空间分辨率下采样至 224 × 224）。这为我们提供了一组 2D DINO 特征的时序序列 $\mathcal{F}_{2D} \in \mathbb{R}^{T_1 \times 16 \times 16 \times 384}$。

无参数反投影. 遵循 [12, 13]，我们首先在规范帧中定义一个三维点的体素网格，并将这些点投影到每个输入帧的像素空间。然后通过对 2D DINO 特征图进行双线性插值来获得每个点的特征编码。这产生了一系列 3D DINO 特征体积，随后通过时间聚合为单个 3D 特征体积 $\mathcal{F}_{3D} \in \mathbb{R}^{Z \times Y \times X \times 384}$，其中 $Z = X = 96, Y = 32$ 是我们 Y 轴向上规范帧中体素的空间分辨率。我们简单地使用时间上的平均池化作为时间聚合方法。

鸟瞰视图投影. 直接在 3D 特征空间中进行推理是昂贵的，并且在许多情况下效果不佳（如后文所述）。因此，我们通过“挤压”上轴（Y 轴），将上述获得的 3D 特征体积投影到“鸟瞰图”（BEV）视图，遵循 [12, 13]。这个挤压过程使用一个多层感知器（MLP）来将展平后的上通道维度（$384 \times Y$）投影为相同大小的通道维度（384）。完成这一步后，我们得到了一个 BEV 特征图 $\mathcal{F}_{\text{BEV}} \in \mathbb{R}^{Z \times X \times 384}$。

BEV **网络**. $\mathcal{F}_{\text{BEV}}$ 是我们进行大部分计算的压缩特征嵌入。遵循之前的设计 [12, 13, 43]，我们的 BEV Net 由 11 个顺序的 BEV_模块模块组成，每个模块应用一个 2D 卷积、LayerNorm 和带有 GELU 激活函数的 MLP。隐藏维度从 384 开始，在中间翻倍两次，最终特征维度达到 1540，同时将空间维度减少到 3×3。

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

轨迹预测. 我们首先对来自 BEV Net 的特征进行空间平均池化，然后使用具有 LayerNorm 和 GELU 激活函数的 3 层 MLP 来获取预测的未来头部姿态序列 $\hat{\mathcal{H}}_{T_1+1:T_1+T_2}$。

损失函数。我们在转换和旋转上使用结合的 L1 损失来监督模型，遵循 [26]：

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

其中 R 是从 6D 旋转表示转换而来的旋转矩阵，I 是单位矩阵。在我们的实验中我们使用了 $\lambda_{\text{trans}} = \lambda_{\text{rot}} = 1$。

3.3. 实现细节

头中心规范框架. 由于我们没有将过去的头部姿态输入模型（它们仅在反投影过程中使用），我们需要定义一个相对于当前头部姿态的规范框架 $\boldsymbol{h}_{T_1}$。这样的规范框架已在先前的工作中被广泛采用 [11, 26, 33, 49, 65]。遵循 [33]，我们也定义我们的框架与地面平行并朝前，但中心位于头部。这使得模型能够在面向当前行进方向的空间中运行，并以相对的方式预测未来的头部姿态。

训练细节. 我们使用 AdamW [28] 优化器，并将 0.05 的权重衰减应用于所有模型参数，但不包括偏置。我们训练我们的模型 700k 步，并使用 OneCycle 学习率调度器 [55]，采用线性退火策略，并将开始百分比设置为 0.05。我们使用的批次大小为 4。在单个 NVIDIA RTX A6000 GPU 上，训练大约需要 4 天完成。

4. 阿里亚导航数据集（AND）

正如在 Sec. 2 中讨论的，据我们所知没有适用于我们的任务的数据集。因此我们设计了一个数据收集管道来收集自己的数据集，并将其命名为 Aria 导航数据集（AND）。接下来我们将介绍关键的数据收集步骤和数据集统计信息。

4.1. 数据收集管道

硬件 . 我们的数据采集硬件仅由一对 Project Aria 眼镜 [7] 组成，与之前部署自建传感器套件 [36, 63] 或遥操作机器人 [17] 的工作相比，它具有轻便、不显眼、成本低廉和易于设置等几个关键优势。

记录过程 。项目 Aria 配备了一个名为 Aria Studio 的移动应用程序，该应用程序允许用户在每次录音会话前通过与移动应用进行一次交互来轻松记录数据。该应用程序提供了对已录制数据模式的选择，在我们的流程中，我们激活了 RGB、SLAM（两个黑白摄像头）和眼动追踪摄像头以及 IMU 传感器、气压计和 GPS。所有摄像头均以 20fps 的帧率运行。在每次录音会话之前，人类受试者会选择此保存的录音配置文件并在四处走动时开始录制，没有预先定义的指令或脚本。为了捕捉一致的信息收集行为，我们指示受试者遵循仔细的导航行为，e.g. 始终在穿越道路前检查过往车辆。

数据处理 . 原始记录的数据以一种名为 VRS [†] 的压缩格式存在。我们运行 Aria Machine Perception Services [‡] 来获取处理过的数据模式，包括 6D 头部姿态轨迹和场景点云。由于鱼眼相机的原因，原始的 RGB 帧发生了畸变，所以我们对其进行去畸变以作为模型的输入。原始序列进一步被分割成长度为 (T_1+T_2) 的片段，采用滑动窗口的方式进行，步幅和膨胀系数均为 6 帧。在每秒 20 帧的情况下，每个片段覆盖了 $(8+8-1)\times 6/20 = 4.5$ 秒的时间。在 SLAM 过程中，会对动态物体上的点进行过滤。我们还会对重建的点云应用另一个过滤过程以去除噪声点。

隐私 。我们已采取措施遵循 Project Aria 研究指南。我们也使用最先进的去标识算法 [48] 来模糊所有视频中的面部。

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

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

4.2. 数据集统计

位置 。由于我们希望捕捉现实世界的导航场景，其中人类需要避免与静态和动态障碍物发生碰撞，因此我们在室内和室外交通密集的地方选择了多样化的地点。我们从大学校园、城市市中心、公园等地选取了 18 个人口稠密的地点。这些地点中许多都非常广阔，为收集的数据提供了极大的多样性。我们特别选择在通常出现交通高峰的时间进行数据记录，e.g. 下课后。我们也多样化了时间段分布。

数据规模 。我们记录了大约 4 小时的数据，在处理后得到了 274k 张 RGB 帧和 36k 个片段。

5. 实验结果

我们将 LookOut 与基线定量比较在 Sec. 5.1.1。在 Sec. 5.1.2 中，我们消融了我们的关键设计选择。然后我们在 Sec. 5.2 中展示了定性评估结果，这些结果显示了我们的模型在现实世界的导航场景中学到的多样化行为。最后，我们调查了来自我们模型的失败案例，并讨论了限制在 Sec. 5.3 中的情况。所有结果都是在一个训练期间未见过环境的保留集上获得的。我们鼓励读者查看我们的项目网页，其中包含通过连续滚动我们的模型给定每个传入帧而获得的视频版本的 Fig. 3，类似于其在实践中运行的方式。

5.1. 定量评估

指标 。我们首先通过与训练时相同的误差函数评估头部姿态预测的准确性，即平移（L_1_trans）和旋转（L_1_旋转变换）上的 L1 损失 i.e. 。为了测量与环境的碰撞情况，我们分别为静态（Col_stt_k）和动态（Col_dyn_k）障碍物定义了一个非碰撞分数。得分衡量的是预测值中至少距离最近障碍物 k 厘米的百分比。对于静态障碍物，我们测量预测头部平移与 SLAM 重建点云之间的最短距离。而对于动态障碍物，我们首先使用单目度量深度估计方法 Depth Pro [1] 为数据集中的每一帧估算一个深度图，然后使用 DINOv2 + Mask2Former 分割头 [38] 获取每帧的语义分割掩码。我们随后将所有标记为“人”的像素中估计度量深度值的最小值作为最短距离。Col_*_平均是所有 $k \in \{15, 25, 35\}$ 的平均值。请注意，非碰撞得分是避碰的一个粗略替代指标，我们也报告了真实序列（地面真相）的值以供参考。

方法	L_1_trans ↓	L_1_rot ↓	Col_stt_15 ↑	Col_dyn_15 ↑	Col_stt_25 ↑	Col_dyn_25 ↑	Col_stt_35 ↑	Col_dy
Const_Vel	0.41	0.77	85.5	91.2	80.0	81.3	74.1	73
Lin_Ext	0.45	1.21	86.5	92.3	77.6	82.1	73.3	72
EgoCast [8]	0.34	0.63	90.5	94.6	84.6	86.2	77.4	77
Ours	0.17	0.16	91.3	97.2	85.6	90.3	79.9	83
GT	0	0	92.7	97.7	88.9	93.0	83.6	85
A*+Lin_Ext	0.24	1.21	98.8	82.4	100.0	76.5	100.0	61
Ours (+goal)	0.11	0.15	91.7	97.2	86.3	91.4	82.0	84

表 1. **与基线的比较**。我们的模型在轨迹预测和碰撞避免方面都优于可比较的方法。

5.1.1. 与基线比较

基线 . 由于我们研究的是一个新颖的问题，没有可以直接比较的先前工作。正如所提到的，与我们的工作最接近的研究是 EgoNav [63] 和 EgoCast [8]。EgoNav 是一篇未发布代码的 Arxiv 预印本。因此我们将其适应到我们的设定中。EgoCast 的核心部分是一个基于变压器的预测模块，它可以根据过去的全身姿态和可选的以自我为中心的视频来预测未来的 3D 全身姿态。为了处理在实际操作中往往无法获得全身姿态的问题，该研究进一步实现了一个估计模块，从过去的 6D 头部姿态和以自我为中心的视频中估算当前帧的全身姿态。这两个阶段都是通过 3D 全身姿态进行监督的，而我们的数据集中并没有这些。因此我们重新利用了预测模块来接收过去的头部姿态（而非全身姿态）和以自我为中心的视频，并且也预测未来的头部姿态。然后移除了估计模块。我们在与我们的模型相同的训练分割上对其进行训练。

我们还实现了以下基于过去头部姿势的基线模型：(1) 恒定速度（恒定速度）使用从最后两个输入步骤计算出的线性和角速度来推算未来的头部平移和旋转，(2) 线性外推（线_扩展)），该模型对过去的平移和旋转序列拟合一个线性回归模型并预测未来的情况，以及 (3) A*+ 线性外推（A*+林_扩展）使用线性外推进行旋转，但对于平移则实现了一个 A* 算法。具体来说，我们通过将 SLAM 重建的点云转换为占用栅格来离散化空间，并使用与我们的模型相同的空间分辨率。然后我们将未来 $T_1 + T_2$ 步骤中的最后一个地面真实头部平移作为目标。我们也定义了一个最大速度，该速度大致匹配人类运动能力。我们使用一种变体模型，该模型也将这样的目标位置作为输入（直接连接到最终的 MLP）以公平地与这个基线进行比较。

分析 . 比较结果报告在 Tab. 1 中。在没有目标设定的情况下（顶部），我们的模型在所有指标上都取得了最佳性能，能够预测准确的头部姿态并可靠地避免与静态和动态障碍物发生碰撞。当提供目标时，A*+林_扩展对静态障碍物几乎达到了完美的非碰撞得分，因为它们通过场景占用率进行了显式建模（其中每个体素网格表示大约 600 厘米 ^{3} 的空间），并且搜索算法基本上保证了绕过已占据区域的路径。然而，由于动态障碍物没有在点云中表示，这个基线在避免动态障碍物方面表现不佳。

5.1.2. 消融研究

我们消融输入数据模态和关键模型设计，并在 Tab. 2 中总结结果。

多模态支持 。我们的模型可以很容易地扩展以集成额外的传感器模式，e.g. 深度和点云。具体来说，我们首先将深度转换为点云，然后将其转化为占用体素并与 3D DINO 体积 $\mathcal{F}_{3D}$ 连接。如预期的那样，结合这些直接与障碍物接近相关的模式提高了非碰撞指标，这与之前的研究结果一致 [40]。特别是使用深度有助于避免动态障碍物，因为 SLAM 重建的点云仅包含静态对象。

模型设计 . 我们首先通过消融一种不经过 DINO 而直接反投影原始 RGB 帧的变体来验证 DINO 特征编码的有效性 (无 DINO)。如所示，DINO 特征由于其强大的语义特征编码能力显著提升了模型的表现。接下来我们检查中间三维特征空间的影响，为此我们消融了一种在 2D DINO 特征上进行时间池化的变体 $\mathcal{F}_{2D}$(仅限二维)。我们可以看到，三维特征空间通过赋予特征

方法	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

表 2. **消融研究**。每个设计选择都有助于性能。

一个明确的几何概念提升了性能。最后，我们通过对比直接对 $\mathcal{F}_{3D}$ 应用 3D 卷积来研究 BEV 投影是否有益 (3D 卷积)。这种变体的表现与我们的模型相当，但由于使用了 3D 卷积，在计算上更为昂贵。

5.2. 定性评估

多样化的模型行为．我们通过视觉检查来评估我们的训练模型是否表现出期望的行为。具体来说，我们关注的是（1）我们的模型预测的轨迹是否避免了与静态和动态障碍物发生碰撞，(2) 我们的模型是否学会了类似人类的信息收集行为，以及（3）我们的模型在何处失败。为此，我们在二维空间中可视化了我们模型的预测结果和真实值，并将它们叠加到图像观察之上。我们在 Fig. 3 中展示了这些可视化的几个样本，在项目网页上还有更多展示。可以看出，我们的模型预测了绕过静态和动态障碍物的安全路径。我们的模型还学会了在训练数据中人类所表现出的信息收集行为，即它会预测头部旋转以检查可能有用的导航信息（e.g. 道路状况）。我们还观察到了模型的其他有趣行为。在第一个和第五个例子中，当没有容易通行的路径时，模型学会了等待。在第三个例子中，模型根据新观察到的视觉线索调整其预测结果（人出现后，预测路径从中间向右移动）。

BEV **可视化**．我们还展示了从 BEV 转换后的可视化翻译 Fig. 4。我们将轨迹叠加在静态环境的 BEV 表示之上，这是通过将场景点云转换为占用栅格图，然后再转换为高度图获得的。高度图存储每个像素沿向上轴的所有被占据网格的最大高度。从这些可视化中可以看到，我们模型预测的轨迹满足各种场景下的环境约束。

5.3. 失败案例和限制条件

我们识别出模型的失败案例，并在 Fig. 5 中提供了可视化。我们模型的一个主要限制在于缺乏生成建模能力，因此当未来是多模态可能时可能会遇到困难。在 Fig. 5 的第一个示例中，可以通过向左或向右走来避免与迎面而来的行人发生碰撞，在这种情况下我们的模型应该回归到这些多种可能性的平均值。只有在这种情况下的人类主体明显走向他们的右侧时，我们的模型才能回归到一种可能的未来。因此下一步是利用生成模型学习这样的多模态分布，例如扩散模型 [15, 57] 。在第二个案例中，人类主体低头检查中间时间步中的铁轨位置以避免绊倒。然而，由于在我们的训练集中从未出现过铁轨，我们的模型并未做出此类预测。扩展我们的训练集以包含更多样化的场景将是一个解决这个问题的有希望的方法。

6. 结论

在本文中，我们通过一系列贡献迈向了可部署于现实世界的人形导航策略。首先，我们在存在静态和动态障碍物的情况下，提出了一项新的任务，即从过去的以自我为中心的视频预测未来的 6D 头部姿态轨迹。这项任务定义使得模型不仅能够学习规划无碰撞路径，还能够学习类似人类的信息收集行为。其次，我们提出了一种利用预训练的 DINO 特征编码器和一种无需参数的反投影策略来有效解决该任务的模型。接下来，我们设计了一个数据采集管道，仅使用一对 Project Aria 眼镜作为数据捕获设备。此管道易于扩展，并允许我们轻松地收集一个 4 小时的真实世界导航数据集。我们的数据集涵盖 18 个地方，交通多样且密集，为社区提供了一种有价值的资源。通过广泛的实验，我们展示了我们的模型学习了对现实世界导航任务有用的多样化行为，并在所有指标上超越了基线模型。最后，我们讨论了我们模型的失败案例和局限性以及未来工作的方向。

致谢

本工作得到了 ARL 资助号 W911NF-21-2-0104 的支持，以及 Vannevar Bush Faculty Fellowship、HAI 和 Meta Reality Labs 的资助。

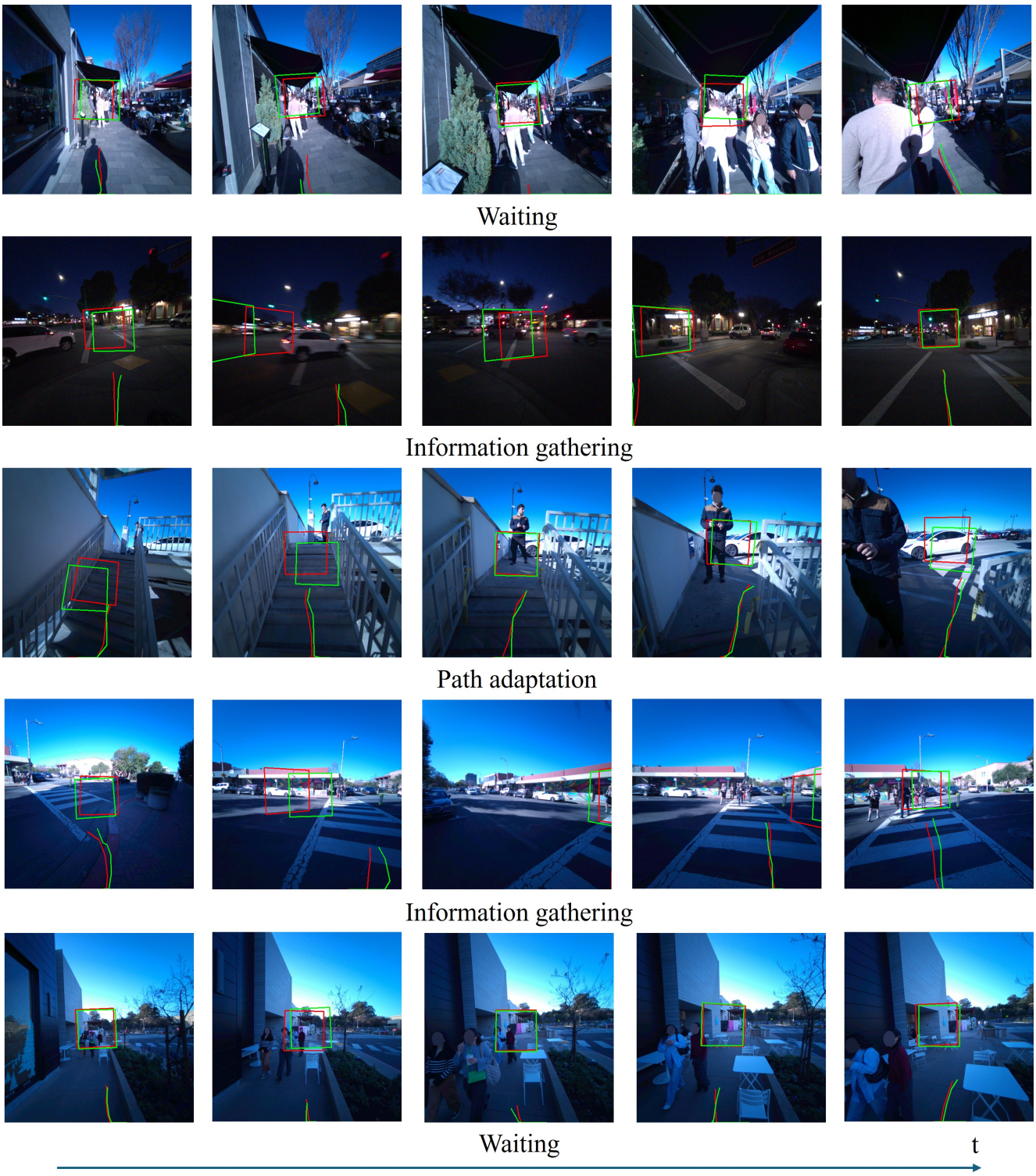

图 3. **模型行为的可视化**. 我们从保留集提供了五个示例，包含模型预测（红色）和真实值（绿色）。对于每个示例，展示了五帧图像并在下方附有描述模型行为的文本。通过将值投影到地面再投影到图像平面上获得翻译可视化（曲线）。旋转（方块）是通过将视锥体投影到图像平面上进行可视化的。我们显示完整的未来序列用于翻译，而仅显示下一个步骤用于旋转以便于清晰理解。

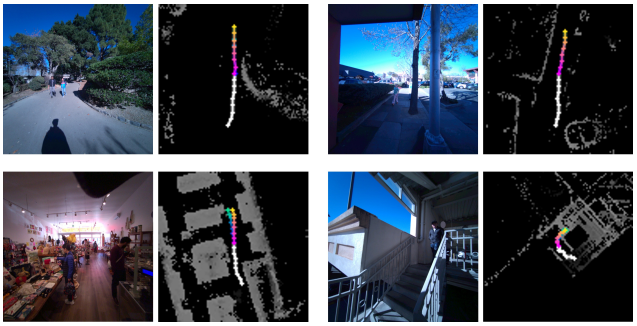

图 4. BEV **可视化**. 我们展示四个示例，每个示例左侧是一个采样的 RGB 帧，右侧是轨迹的 BEV 可视化。白色曲线表示过去，蓝绿色表示真实的未来，而粉橙色代表预测的未来。颜色编码描绘了时间进程的顺序。这些轨迹叠加在场景点云的 BEV 表示上以进行可视化。请注意，这里仅显示了轨迹的平移分量。

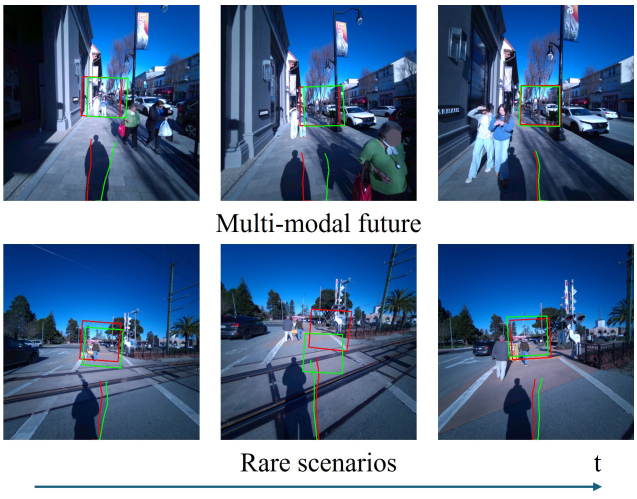

图 5. **失败案例**。我们展示了来自保留测试集中的两个示例，并在每个示例下方描述了失败原因。

参考文献

[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 multimodal 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, 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. 7

[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, 4, 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. 7

[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, 6

[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