Extended Abstract

Motivation For autonomous mobile robots to navigate highly dynamic, human-crowded environments safely, they must respect implicit social etiquette, personal spacing, and pedestrian flow. Abstracted and hard-coded geometric rules cannot properly capture these complex dynamics, which can lead to issues like collisions and injury. It can also cause planners to fail to find valid paths, making it freeze in place and not progress. Reinforcement Learning (RL) has been applied to many similar scenarios and planning/navigations tasks however, online RL is unsafe to train around humans, and offline RL can struggle with reward engineering instability in multi-agent settings and implicitly learning the reward in a social scene can be unclear. We propose transforming social navigation into a visual, Vision-to-Action (V2A) image-to-image translation task using Imitation Learning to learn social planning directly from expert demonstrations.

Method We present a Vision-Based Goal-Conditioned Behavioral Cloning (GCBC) framework. We project raw 360-degree panoptic semantic masks and LiDAR data into a flat, egocentric Bird's-Eye View (BEV) grid. Stacking these grids to provide temporal history, and concatenating a goal mask formulates a spatial state representation of the surrounding environment. A Res-UNet architecture is then used to predict a continuous multi-step Spatial Action Map, representing the probability distribution of future waypoints. This avoids the possible mathematical instability of predicting continuous control velocities using Convolutional Neural Networks (CNNs)

Implementation We conducted an ablation study using the JRDB-PanoTrack dataset over 100 training epochs. We evaluated three distinct paradigms: (1) A Baseline semantic representation, (2) A 1-Hot Encoded Multichannel representation expanding the state to 297 channels, and (3) A "Social Cloning" paradigm. In Social Cloning, we utilized the exact same single-channel architecture as the baseline, but shifted the egocentric frame from the robot to background pedestrians. To accurately mimic the JackRabbot mobile robot, we cleared a 0.25m radius around the pedestrian. This transformed them into expert human demonstrators and expanded our training dataset from 847 frames to 5,148 frames.

Results The baseline model achieved reasonable convergence (Val ADE of 0.63m) but struggled with complex social flows, resulting in a Robot Social Collision Rate (SCR) of $1.47\% \pm 0.10\%$. The Multichannel (1-Hot Encoded) approach improved the ADE to 0.3875m, but caused a severe computational bottleneck, taking 16 times as long to train due to the added complexity. The Social Cloning approach achieved the best overall performance, yielding a Validation Robot ADE of 0.3475m and a low Robot SCR of $0.17\% \pm 0.05\%$.

Discussion Directly predicting Spatial Action Maps with Res-UNet preserves the spatial understanding of the environment. Social Cloning proved that learning from human-to-human interactions transfers well to robot-to-human navigation, leading to the smoothest trajectories and best results. 1-Hot Encoding provided more explicit semantic distinction, but the large increase in channel depth caused a computational bottleneck that made it less scalable than the Social Cloning data augmentation approach.

Conclusion By combining 360BEV semantic mapping with Goal-Conditioned Behavioral Cloning and introducing Social Cloning, we developed a highly stable and safe path planner. The framework successfully models complex human movement without the need for explicitly engineered spatial reward functions, achieving very high collision avoidance on hold-out data.

Goal Conditioned Behavior Cloning for Robot Social Navigation

Mete Gumusayak
Department of Computer Science
Stanford University
mete1@stanford.edu

Abstract

Navigating human-crowded spaces requires autonomous robots to understand social etiquette, personal spacing, collision avoidance and pedestrian flow. We propose a Vision-to-Action (V2A) framework utilizing Goal-Conditioned Behavioral Cloning (GCBC) to predict optimal spatial trajectories from semantic data. By projecting 360-degree panoptic labels and LiDAR data into an egocentric Bird's-Eye View (BEV), we allow a Res-UNet to predict continuous spatial action heatmaps. This method was used to avoid potential instability when predicting continuous control velocities with CNNs. To address the sample inefficiency of Behavioral Cloning, we also introduce a data augmentation technique called *Social Cloning*. This technique extracts moving background pedestrians, removes their physical footprint from the state to mimic the physical robot (which is not in the scene), and frames them as expert ego-demonstrators, yielding a large increase in training data. Our ablation study demonstrates that Social Cloning significantly outperforms baseline representations and high-dimensional 1-Hot Encoded representations. It achieves an Average Displacement Error (ADE) of 0.3475m and reduces the Social Collision Rate to 0.17%.

1 Introduction

Integrating autonomous mobile robots into human-populated spaces—such as cafes, university campuses, and walkways—requires a sophisticated understanding of environmental dynamics. Traditional navigation stacks rely heavily on costmaps and hard-coded geometric rules. However, these analytical methods cannot properly capture the complex dynamics of human crowds, which can lead to issues like collisions and injury. It can also cause planners to fail to find valid paths, making the robot freeze in place and not progress, known as the "freezing robot problem."

Reinforcement learning (RL) has been applied to many similar scenarios and planning/navigation tasks. However, online RL is fundamentally unsafe to train around humans because the agent must explore and make mistakes to learn. Offline RL removes this safety risk by learning from static datasets, but it can struggle with reward engineering instability in complex settings. Implicitly learning the reward in a dynamic social scene can be unclear, and designing a reward function that accurately penalizes socially incorrect actions without causing reward hacking is difficult.

We propose transforming social navigation into a visual, Vision-to-Action (V2A) image-to-image translation task using Imitation Learning to learn social planning directly from expert demonstrations. We present a Vision-Based Goal-Conditioned Behavioral Cloning (GCBC) framework. By projecting 360-degree panoptic segmentation masks into a Bird's-Eye View (BEV) grid, we use the inductive biases of Convolutional Neural Networks (CNNs). Instead of predicting continuous control vectors like linear and angular velocities, our model outputs a Spatial Action Map, a probability heatmap of future waypoints. To overcome the sample inefficiency associated with Behavioral Cloning,

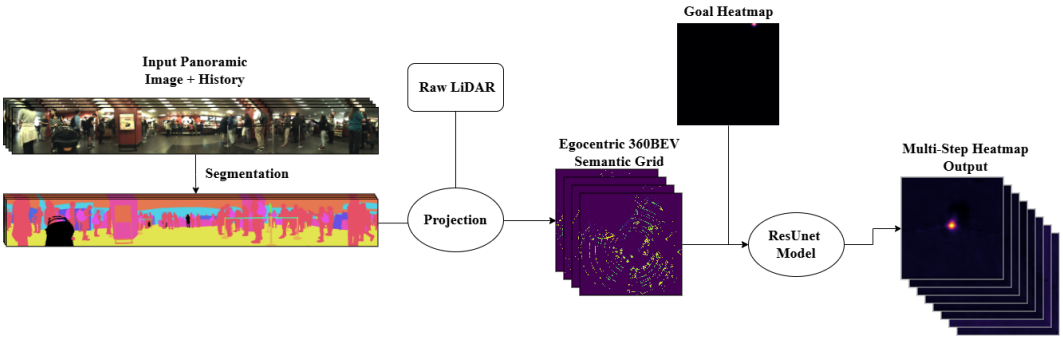

Figure 1: Our Vision-to-Action (V2A) Pipeline: 360-degree Panoptic Labels and LiDAR point clouds are projected into an Egocentric BEV grid. A Res-UNet processes the ego-motion-compensated stacked temporal history alongside a goal mask to output a Multi-Step Spatial Action Heatmap.

we introduce *Social Cloning*. This methodology transforms background pedestrians into expert demonstrators, dramatically increasing the volume of high-quality, socially compliant training data without requiring additional robot deployment.

2 Related Work

End-to-End vs. Modular Navigation Traditional navigation pipelines separate perception, mapping, and planning into distinct modules. These systems are highly interpretable however, errors in early stages (like an imperfect bounding box around a pedestrian) pass into later modules, leading to catastrophic failures. End-to-end learning methods attempt to map raw RGB video directly to control actions. Datasets such as EgoWalk(4) and JRDB-PanoTrack (1) provide human navigation data to train these models. However, learning control directly from RGB is highly sample-inefficient due to varying lighting, textures, and visual noise present in all environments. In this work, we utilize the panoptic semantic masks from JRDB-PanoTrack as a mid-level latent representation. This allows the model to extract general, navigation-relevant features while ignoring irrelevant visual textures.

Learning-Based Social Navigation In learning-based navigation, Goal-Conditioned Behavioral Cloning (GCBC) is used to imitate expert trajectories without manual reward engineering. Offline RL like Implicit Q-Learning is used to improve suboptimal demonstrations however, GCBC is a stable and effective method when an abundance of high-quality expert data is available. Our work focuses on scaling the availability of this expert data through perspective transformations rather than relying on temporal-difference learning, which can be unstable in environments with unpredictable human agents.

Bird's-Eye View and Action Representations The choice of state and action representation is very important to the stability of the learned policy. Predicting raw 1D continuous vectors (velocity and yaw) from a compressed image embedding forces the network to implicitly learn the camera's spatial geometry, which can result in erratic driving. Representing actions as Spatial Action Maps (3) allows models to utilize fully convolutional architectures like UNet. This architectural combination preserves the spatial geometry of the BEV grid from input to output. Frameworks like 360BEV (2) have demonstrated the efficacy of projecting panoramic data into 3D. However, prior applications of 360BEV focused primarily on static indoor mapping, and we adapt this method for dynamic, multi-agent trajectory forecasting.

3 Methodology

Our approach views social navigation as a visual, Goal-Conditioned Behavioral Cloning (GCBC) problem. The pipeline has four primary stages: (1) transforming raw panoramic and LiDAR sensor data into a structured spatial representation, (2) encoding temporal history and navigational goals, (3) augmenting the dataset via Social Cloning, and (4) predicting multi-step trajectories using a fully convolutional network and spatial loss functions.

3.1 Ego-Centric BEV Projection

To create an accurate top-down representation of the environment, we adapt the mapping paradigm from 360BEV (2), which uses RGB-D Pano images. We chose this method because relying on a flat-ground Inverse Perspective Mapping (IPM) causes severe distortion, particularly elongating vertical objects like pedestrians. To implement this, we utilize the JackRabbot's dual (upper and lower) Velodyne LiDAR point clouds combined with its RGB Pano images.

Let a 3D point from the LiDAR sensor be denoted as $\mathbf{P}_L = (X, Y, Z)$. We first transform this point into the panoramic camera's coordinate frame using the extrinsic calibration matrix $T_{L\to C}$:

$$\mathbf{P}_C = T_{L\to C}\mathbf{P}_L \tag{1}$$

We then apply a projection to map $\mathbf{P}_C$ to a 2D pixel coordinate (u, v) on the stitched 360-degree panoptic mask. The semantic class ID at (u, v) is assigned to the 3D point $\mathbf{P}_L$.

Once the point cloud is semantically color-coded, we rasterize the points onto a 2D Bird's-Eye View (BEV) grid. We define the BEV spatial dimensions as 128×128 pixels, capturing a local area of $15\text{m} \times 15\text{m}$ around the robot, yielding a spatial resolution of0.12 meters per pixel. The Z-axis (height) is removed, preserving the highest point that can be interacted with by the robot(< 1 m), generating a semantic footprint map without flat-ground distortion.

3.2 Temporal State and Goal Representation

Static frames lack the velocity and directional information necessary for safe social navigation. Our state representation, s_t, uses a stacked temporal history of semantic BEV grids.

Ego-Motion Compensation Stacking frames from time steps $t, t-1, t-2$, and $t-3$ is insufficient, as the robot's own movement causes static obstacles to appear as if they are moving. We use the robot's odometry to compensate for ego-motion. This ensures that the scene aligns with the robot's current pose, meaning only dynamic objects like pedestrians will exhibit displacement across the channels.

Goal Mask Generation To condition the policy, we compute a local waypoint $w_g = (x_g, y_g)$ located several timestemps ahead on the expert's future trajectory. If w_g falls outside the $15\text{m} \times 15\text{m}$ BEV boundaries, we project it onto the bounding edge of the grid. This coordinate is encoded as a 1-channel binary mask, $M_g \in \mathbb{R}^{H\times W}$, where the pixel corresponding to w_g is set to 1, and all others to 0.

Encoding Paradigms We evaluate the trade-off between semantic information and computational efficiency by defining two encoding paradigms for the stacked tensor:

- **Baseline Semantic Representation (1-Channel Per Frame):** Each time step utilizes 1 semantic channel of normalized integer category IDs (e.g., $1/72$ class 1, $2/72$ for class 2). Stacking 4 frames yields a $(4, H, W)$ tensor. Concatenating the Goal Mask M_g results in a total input tensor depth of **5 channels**.
- **Multichannel (1-Hot Encoded):** Treating categorical IDs as continuous scalars can force the network to assume false relationships, thinking lower classes are less important than larger ones. To prevent this, each frame is expanded into 74 explicit binary channels (72 COCO categories + 1 unobserved + 1 unannotated). The semantic tensor shape becomes $(4 \times 74, H, W) = (296, H, W)$. Including M_g, the final input tensor depth to **297 channels**.

3.3 Social Cloning Data Augmentation

Behavioral cloning typically suffers from sample inefficiency because collecting data can be an expensive process. Another problem is that recorded sequences yield only one expert trajectory (the ego-robot's path). We introduce *Social Cloning*. Using the JRDB panoptic tracking labels, we extract the spatial trajectories of every moving background pedestrian.

Let (x_r, y_r, θ_r) denote the robot's global pose and (x_p, y_p, θ_p) denote a pedestrian's global pose. For each tracked pedestrian, we mathematically rotate and translate the entire BEV scene so that the

pedestrian is at the origin $(0, 0)$ facing forward ($\theta = 0$):

$$R(\theta_p) = \begin{bmatrix} \cos(-\theta_p) & -\sin(-\theta_p) \\ \sin(-\theta_p) & \cos(-\theta_p) \end{bmatrix}, \quad \begin{bmatrix} x' \\ y' \end{bmatrix} = R(\theta_p) \begin{bmatrix} x - x_p \\ y - y_p \end{bmatrix} \tag{2}$$

This perspective shift transforms background pedestrians from obstacles into expert ego-demonstrators navigating the crowd. This data is also filtered to only include pedestrians that are tracked for a minimum of 20 seconds and have an average speed of 0.2m/s to only include higher quality data more similar to how we want the robot to move.

Simulating the Robot Footprint: Shifting the ego-frame leaves the pedestrian's own semantic pixels at the center of the transformed image. If left untouched, the network would learn to generate trajectories that pass directly through human-labeled pixels. To mimic the JackRabbot mobile robot, we apply a spatial mask M_{clear} to the center of the transformed BEV:

$$M_{clear}(x, y) = \begin{cases} 0 & \text{if } \sqrt{(x - x_c)^2 + (y - y_c)^2} \leq 0.25\text{m} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

This clears the human demonstrator and enforces the physical footprint of the robot, generating more realistic, safe synthetic training data.

3.4 Res-UNet and Spatial Action Maps

We implement a Res-UNet architecture to predict the Spatial Action Map. The network utilizes a pre-trained ResNet encoder to compress the spatial features, followed by a decoder with skip connections to recover the exact input spatial resolution (128×128). The network does not pool features into a 1D vector to regress raw decimal outputs; instead it keeps the fully convolutional nature of our model, preserving the spatial geometry from input to output.

The network predicts a multi-channel probability heatmap $\hat{Y} \in \mathbb{R}^{f \times H \times W}$, corresponding to a sequence of $f = 8$ future waypoints. During training, the ground truth heatmap Y is constructed by plotting the expert's actual future coordinates at time $t' = t + 15$ onto the BEV grid and applying a 2D Gaussian blur. For an expert waypoint at (x_{gt}, y_{gt}), the ground truth value at spatial location (i, j) is:

$$Y_{i,j} = \exp\left(-\frac{(i - x_{gt})^2 + (j - y_{gt})^2}{2\sigma^2}\right) \tag{4}$$

where σ controls the variance (spatial tolerance) of the target. We use this method because real robots have a spatial tolerance where there are several correct paths as long as they are with a small tolerance, we use $\sigma = 2$px. We optimize the network using a variant of the 2D Gaussian Focal Loss, defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i,j} \begin{cases} (1 - \hat{Y}_{i,j})^{\alpha} \log(\hat{Y}_{i,j}) & \text{if } Y_{i,j} = 1 \\ (1 - Y_{i,j})^{\beta} (\hat{Y}_{i,j})^{\alpha} \log(1 - \hat{Y}_{i,j}) & \text{otherwise} \end{cases} \tag{5}$$

where N is the number of waypoints, and $\alpha = 2$ and $\beta = 4$ are focal hyperparameters that penalize false negatives and reduce the penalty for predictions that are within the safe Gaussian radius surrounding the exact expert coordinate. At inference time, a simple $\text{argmax}(\hat{Y})$ extracts the highest-probability coordinate to pass to the robot's low-level controller.

4 Experimental Setup

We trained our models utilizing the JRDB-PanoTrack dataset over 100 epochs. The networks were optimized using the AdamW optimizer. We conducted an ablation study evaluating three specific configurations:

1. **Baseline (Default):** The 1-channel per frame encoding (5-channel input), trained purely on ego-robot trajectories. The dataset splits were **847** train frames, 131 validation frames, and 158 test frames.
2. **Multichannel (1-Hot):** The 1-Hot Encoded semantic channels (297-channel input), trained purely on ego-robot trajectories. Dataset splits matched the baseline (Train: 847, Val: 131, Test: 158).

3. **Social Cloning:** This paradigm used the same single-channel architecture as the baseline (5-channel input), but was trained on the combined ego-robot and human pedestrian augmented dataset. The dataset splits expanded to **5,148** train frames, 680 validation frames, and 685 test frames.

To evaluate safety and spatial accuracy, we utilize two metrics. The **Average Displacement Error (ADE)** measures the Euclidean distance in meters between the predicted trajectory and the ground truth expert path. The **Robot Social Collision Rate (SCR)** is defined as the percentage of frames where the robot's predicted trajectory passes within a 0.75-meter radius of any person in the environment.

5 Results

5.1 Quantitative Evaluation

The quantitative results of our ablation study are detailed in Table 1.

Table 1: Ablation Study: Performance Comparison Across Paradigms (100 Epochs)

Model Paradigm	Train Frames	Val Robot ADE ↓	Robot SCR (0.75m) ↓
Baseline (Default)	847	0.6300m	$1.47\% \pm 0.10\%$
Multichannel (1-Hot Encoded)	847	0.3875m	$0.48\% \pm 0.05\%$
Social Cloning (Human+Robot)	5,148	**0.3475m**	$\mathbf{0.17\% \pm 0.05\%}$

The Baseline model achieved reasonable path-following capabilities (Val ADE of 0.63m) but struggled significantly with complex pathing and social flow, resulting in a higher collision frequency (SCR of 1.47%).

The Multichannel (1-Hot Encoded) method improved spatial awareness, reducing the ADE to 0.3875m and the SCR to 0.48%. By explicitly separating semantic classes into their own channels, the network was no longer forced to interpolate meaning from scalar values. However, this approach introduced a severe computational bottleneck, detailed below.

The *Social Cloning* paradigm achieved the best overall performance. By keeping the lightweight 5-channel architecture but expanding the dataset to 5,148 frames using human trajectories, the validation Robot ADE dropped to 0.3475m. The model achieved a low Robot Social Collision Rate of $0.17\% \pm 0.05\%$ on the validation set.

5.2 Computational Trade-offs and Bottlenecks

The 1-Hot Encoded Multichannel approach gave better metrics than the baseline, however, this state representation of 297 burdened the pipeline. This caused a severe computational bottleneck, taking **16 times as long to train** due to the added complexity. Specifically, the training throughput dropped to 4 seconds per iteration. In contrast, the Social Cloning approach utilized the default 5-channel architecture, allowing it to process a dataset 6 times larger but in less time compared to the 1HE, making it vastly more scalable.

5.3 Qualitative Analysis

Visual tracking of the predicted f-step trajectories demonstrates that the Res-UNet successfully models complex, real-world expert paths. Figure 2 visualizes the network's internal representations, showing its ability to interpret the semantic BEV, identify the localized goal, and generate a precise spatial action map.

Figure 6 highlights the practical benefits of the Social Cloning augmentation. When tested on a highly dynamic crowded scenario, the baseline model's predicted path becomes erratic as it attempts to navigate around multiple moving agents. The model trained with Social Cloning has a much smoother, deliberate trajectory. It demonstrates better performance in both going straight and turning. While it slightly strays from the exact expert path, it rarely gets close to people, effectively routing around a crowd without invading personal space.

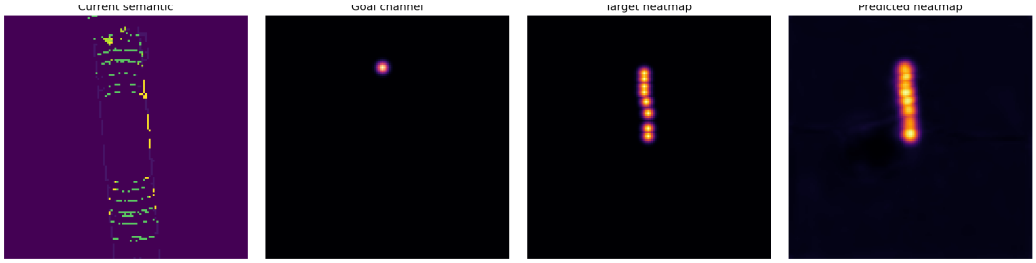

Figure 2: Model Inference Pipeline: (Left to Right) 1. Current Semantic BEV, 2. Target Goal Heatmap, 3. Ground Truth Trajectory Heatmap, 4. Predicted Spatial Action Map output by the Res-UNet.

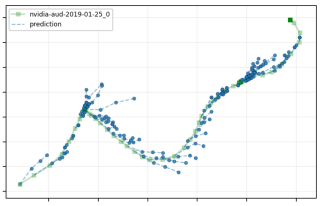

Figure 3: * Baseline Model

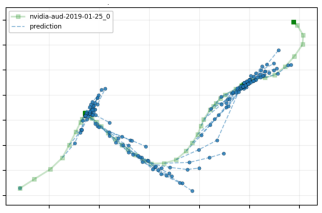

Figure 4: * Social Cloning Model

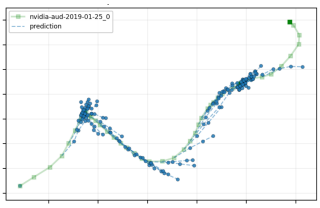

Figure 5: * 1HE Model

Figure 6: Side-by-side comparison of trajectory predictions in a crowded scenario. The Baseline model (left) exhibits a more erratic, unstable path. The Social Cloning model (middle) demonstrates better performance in straight sections and turning maneuvers. It slightly strays from the center of the expert path, but it rarely breaks the 0.75m threshold of pedestrians, reflected in its 0.17% SCR. The 1HE model (right) also performs visually better, staying closer to the path compared to the baseline.

6 Discussion

Directly predicting Spatial Action Maps with Res-UNet effectively preserves the spatial understanding of the environment. The integration of Social Cloning proved that learning from human-to-human interactions transfers well to robot-to-human navigation, leading to the smoothest trajectories and best results.

Interestingly, while the Social Cloning model achieved a Validation Robot ADE of 0.3475m, its Validation *Human* ADE (when evaluated on the hold-out human trajectories) was 1.4307m. This shows that predicting highly erratic, unconstrained human-to-human flow is much more difficult than predicting the slightly more conservative, goal-directed path of the ego-robot.

One observed limitation of this framework is navigating complex maneuvers. The model occasionally struggles with sharp, self-wrapping turns where heatmap activations become less predictable and diffuse. By rasterizing 3D point clouds into a 2D BEV map, we lose vertical dimensionality (e.g., overhangs and tables), which can distort the navigable space if a sensor is mounted low to the ground. We also have less information about the environment because it is sparse data of the environment from the LiDAR

7 Conclusion

By combining 360BEV semantic mapping with Goal-Conditioned Behavioral Cloning and introducing Social Cloning, we developed a highly stable and safe path planner. The framework successfully models complex human movement without the need for explicitly engineered spatial reward functions, achieving very high collision avoidance on hold-out data. Future work will explore integrating Transformer-based architectures to better process multi-channel spatial data without the massive computational bottlenecks seen in standard convolutions.

8 Team Contributions

- **Mete Gumusayak (Individual Project):** I completed all aspects of this research, including data preprocessing, LiDAR BEV projection math, model architecture implementation, dataset augmentation , and quantitative evaluation.
- **AI Tools Disclosure:** I utilized Google Gemini to generate boilerplate code for data loading, dataset exploration, and plotting visualizations. I utilized Anthropic Claude to assist with debugging projection logic and to parallelize workflows to support multiple workers during the heavy data extraction phase. All core architectural decisions, ablations, and analyses were developed independently.

Changes from Proposal In our initial proposal, we outlined a plan to implement Goal-Conditioned Implicit Q-Learning (GC-IQL) as a value-based offline RL follow-up to our GCBC baseline. However, after successfully implementing the *Social Cloning* paradigm, our purely supervised GCBC approach achieved a low collision rate (0.17% SCR). We concluded that because our data augmentation provided a large amount of high-quality expert demonstrations, the architectural simplicity and training stability of Behavioral Cloning outweighed the need for complex temporal-difference RL. We leave value-based offline methods for future work.

References

[1] Duy Tho Le, Chenhui Gou, Stavya Datta, Hengcan Shi, Ian Reid, Jianfei Cai, and Hamid Rezatofighi. *JRDB-PanoTrack: An Open-world Panoptic Segmentation and Tracking Robotic Dataset in Crowded Human Environments*. arXiv preprint arXiv:2404.01686, 2024.

[2] Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, and Rainer Stiefelhagen. *360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View*. arXiv preprint arXiv:2303.11910, 2023.

[3] Jimmy Wu, Xingyuan Sun, Andy Wang, Ian Miller, Abhinav Gupta, and Marc Toussaint. *Spatial Action Maps for Mobile Manipulation*. Proceedings of Robotics: Science and Systems (RSS), 2020.

[4] Timur Akhtyamov, Mohamad Al Mdfaa, Javier Antonio Ramirez Benavides, Arthur Nigmatzyanov, Sergey Bakulin, German Devchich, Denis Fatykhov, Diego Ruiz Salinas, Alexander Mazurov, Kristina Zipa, Malik Mohrat, Pavel Kolesnik, Ivan Sosin, and Gonzalo Ferrer. *EgoWalk: A Multimodal Dataset for Robot Navigation in the Wild*. arXiv preprint arXiv:2505.21282, 2026.