Egocentric Scene-aware Human Trajectory Prediction

Weizhuo Wang[1], C. Karen Liu[1], and Monroe Kennedy III[1]

Stanford University, Stanford CA 94305, USA

Abstract. Wearable collaborative robots stand to assist human wearers who need fall prevention assistance or wear exoskeletons. Such a robot needs to be able to predict the ego motion of the wearer based on egocentric vision and the surrounding scene. In this work, we leveraged body-mounted cameras and sensors to anticipate the trajectory of human wearers through complex surroundings. To facilitate research in ego-motion prediction, we have collected a comprehensive walking scene navigation dataset centered on the user's perspective. We present a method to predict human motion conditioning on the surrounding static scene. Our method leverages a diffusion model to produce a distribution of potential future trajectories, taking into account the user's observation of the environment. We introduce a compact representation to encode the user's visual memory of the surroundings, as well as an efficient sample-generating technique to speed up real-time inference of a diffusion model. We ablate our model and compare it to baselines, and results show that our model outperforms existing methods on key metrics of collision avoidance and trajectory mode coverage.

Keywords: Egocentric Trajectory Prediction · Diffusion Model · Scene Understanding

1 Introduction

The integration of autonomous systems into human-centric environments, particularly through collaborative robotics, necessitates acute awareness and prediction of human motion. These systems, whether external like autonomous vehicles and mobile robots or wearable like exoskeletons, require precise motion prediction to ensure safety and enhance collaboration. External systems predict human trajectories from a distance to avoid collisions or facilitate physical collaboration, whereas wearable systems, mounted directly on users, rely on an egocentric perspective to predict the wearer's motion for functional response.

Predicting motion from an egocentric perspective involves deciphering the complex relationship between past movements, task semantics, and potential future actions. This task is complicated by the need to account for multi-modal outcomes, such as which path a person might choose at a divergence. Factors influencing these decisions range from high-level objectives to immediate environmental constraints like obstacles and walkable paths. A generative machine

Fig. 1: Different red ribbon illustrates different possible modes in the scene and the size of the ribbon denotes the likelihood.

learning model can learn to predict the motion of a human by observing human data of individuals walking in different scenes. Such models are improved and more efficient when curated features are provided. Scene semantics can help provide curated features that improve generative model prediction.

In this paper, we propose a generative modeling approach to predict the distribution of future trajectories of a person given the environment semantics observed from the egocentric view, as illustrated in Fig. 1. To address the under-constrained and multi-modal nature of the navigation problem, we employ a diffusion model to predict the future trajectories conditioned on the egocentric observation of the environment and the user's past walking trajectory. We introduce a compact representation of the environment to capture a short history of visual observations from the egocentric perspective. This "visual memory" representation encodes both appearance and semantics information of the world. To achieve real-time inference, we introduce a hybrid generation technique that balances between quality and speed of sample generation from the trained diffusion model. Using our hybrid generation method, we can rapidly sample multiple future trajectories in real time, providing downstream applications a distribution of the possible paths a person might take, rather than predicting a single path.

We evaluate our predictive model using real-world navigation data. We collect a comprehensive walking scene navigation dataset centered on user's perspective in diverse indoor and outdoor scenarios. We show that our model can accurately predict a distribution of future trajectories representing human preference taking scene context into account, outperforming the baseline methods on various evaluation metrics. We also conduct ablation studies to validate the design choices in visual memory representation and hybrid generation technique. The dataset and the trained model will be made publicly available.

2 Related Work

Path prediction of humans is an area of interest across robotic and computer vision applications. In this section, we highlight key works in the domains of autonomous vehicles predicting the motion of pedestrians, the prediction of humans during sports activities, and the prediction of humans in social settings leveraging social context. We also highlight recent advances in diffusion models as a method of predicting trajectories.

For autonomous vehicles (AV), predicting the motion of other vehicles and pedestrians is crucial for the AV to plan safe, informed trajectories [12]. Recent approaches focus on extracting motion prediction from demonstrations [6]. Traditionally algorithms for trajectory prediction relied on sequential prediction methods such as recurrent neural networks (RNNs), LSTMs, or GRUs [22]. However, these models show limitations when confronted with complex, multi-modal sensor data due to the hidden state size and forgetting information in long sequences. Transformer-based models provide a powerful alternative offering benefits in multi-modal sensor streams and the ability to handle long-sequence inter-dependence's [9, 23]. For AVs, it is often convenient to reconstruct the environment from a birds-eye-view (BEV) or 2D occupancy grid provided sensory inputs [3, 16, 20]. While such representations are easy to interpret when predictions are limited to a small set of modes or choices, they introduce a "long tail" problem which is particularly impactful when the environment has less structure, presenting a need for research in representations that balance structure and flexibility in trajectory prediction [13]

Trajectory prediction of players during sports is another application with rich existing datasets where there is inherent context related to the strategy of the game being played [2]. These sporting scenarios usually take place where there are few static objects/obstacles, and playing grounds are usually flat, leaving game strategy as the sole motion intent informer [2, 21].

Trajectory prediction of humans in social settings provides motion scenarios where behaviors are conditioned on social norms and settings, and leading models (SocialGAN, SociaLSTM) take these factors into account during prediction [1,4]. Such models have been applied to pedestrian dynamics, showcasing the potential for robots to coexist seamlessly with humans in crowded public spaces. [14] Prominent datasets facilitating these studies include UCY/ETH [7] and the Stanford Drone Dataset [14], which predominantly offer a BEV of pedestrian movement, simplifying the trajectory prediction challenge to a two-dimensional problem. This top-down perspective allows for the tracking of pedestrians through bounding boxes, offering an effective means to study and model pedestrian trajectories in a simplified setting.

The diffusion model, inspired by the physical diffusion process, was originally proposed in 2015 [17] but gained significant traction with the introduction of Denoising Diffusion Probabilistic Models (DDPM) in 2020 [5]. Unlike initial attempts that struggled to predict denoised inputs directly, DDPM introduced an approach that predicts the noise to be removed, marking a breakthrough in model performance. Subsequent developments, such as Latent Diffusion Models

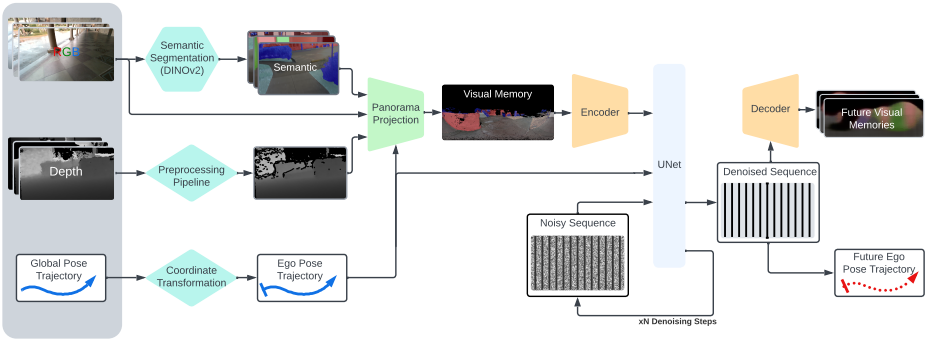

Fig. 2: Overview of the proposed method: We maintain a 5-second buffer of logs that is most relevant to the prediction and organize them into a visual memory frame. All input and output of the prediction module are in the ego-centric frame.

[15] for complex generation tasks, DDIM [18] for accelerated processing, and various guidance and conditioning techniques [8], have expanded the application of diffusion models to image, video, and action generation. These models excel in generating detailed and coherent outputs across a range of generative tasks, proven by many successes in image, video, and even action generation tasks.

3 Method

3.1 Problem Formulation

The goal of this work is to predict the possible paths of a person in a cluttered environment. A trajectory τ is a sequence of 6D poses (translation and orientation) of a person navigating in the 3D world. At each time step t, our model uses the past trajectory $\tau_{1:t}$ to predict likely future trajectories $\tau_{t:t+T}$. In addition, the prediction must be conditioned on the observation of surroundings. The visual observation S encodes the appearance, geometry, and semantics of the environment captured by wearable visual and depth sensors. Our goal is to model the probability distribution of the future trajectories:

$$\hat{\tau}_{t:t+T} \sim p_\theta(\cdot|\tau_{1:t}, S) \tag{1}$$

We learn the model p_θ that best approximates the future trajectory, recognizing that the expected path may be sampled from a multi-modal distribution given the environment. An example of this would be a fork in the path, both paths may be viable, and multiple sampling from the model output distribution should recognize the distributed likelihood.

3.2 Method Overview

Our method takes as input the past trajectory of the person recorded by an Intel Realsense T265 camera and a short history of RGBD images recorded by

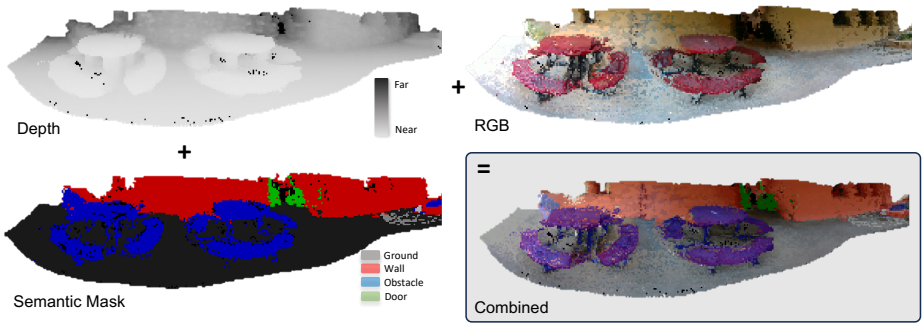

Fig. 3: Channels in Visual Memory: The visual memory integrates frames from various time steps into a single panorama. It consists of depth channel, color channel, and intensity-encoded 8-class semantic channel. 4 channels are shown in the figure.

an outward-facing stereo camera worn by the person and predicts the future trajectories the person might take (Figure 2. The color images are semantically labeled by DINOv2 [10], while the depth images go through a preprocessing pipeline that filters out erroneously filled edges. We transform the past trajectory from a global coordinate frame to an egocentric frame defined by the +Z as opposite to the gravity vector and forward-facing direction as +X. Taking the semantically labeled images, RGB images, and processed depth images, we project and align the images to create a single panorama in the egocentric coordinate frame, referred to as "visual memory". Conditioning on the visual memory and the past ego trajectory, a diffusion model is trained to predict the future trajectory along with encoded visual observations, as auxiliary outputs. The diffusion model starts from a random noise and performs a number of denoising steps to generate the clean prediction sequence. The clean sequence is a concatenation of encoded visual memory and future ego pose trajectory. Finally, we use the VAE decoder to recover the expected future panorama.

3.3 Construction of Visual Memory

Visual memory is defined as an ego-perspective, panoramic view representation of the surroundings. The visual memory is constructed from aligned color, depth, and semantic images. Given the camera's intrinsic and extrinsic parameters, images in different frames can be projected to a single point cloud in the global frame. A distance-based filter is applied to remove points too far away from the current pose. The points are then projected back to the current ego frame to form a coherent representation of all the scene information gathered.

It is important to note that the visual memory representation holds a lot more relevant information than just a single image. As shown in Fig. 4, A single image from a stereo camera only has a narrow FOV pointing directly in front. It fails to capture the objects and paths in the scene that are highly relevant to the prediction, requiring many individual frames to be sent to the predic-

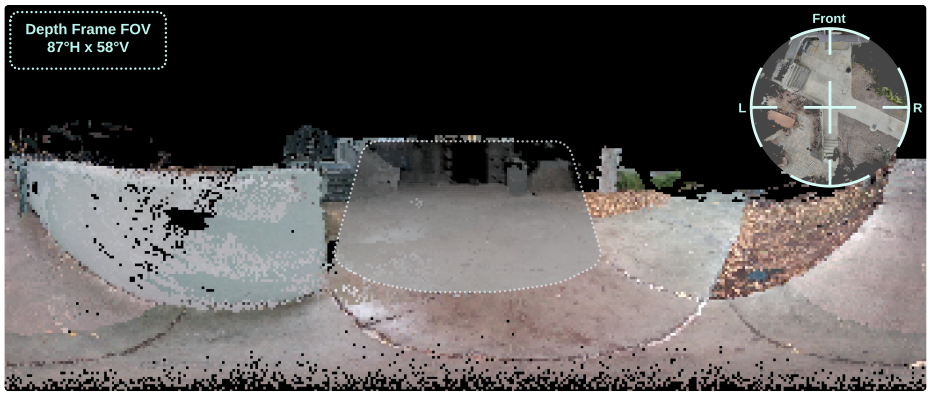

Fig. 4: Comparing depth frame with visual memory: A raw depth frame from stereo camera have only 90 degrees of narrow FOV and often misses important scene information. In the figure, depth frame only sees the open space in front, did not capture the stairs, the right turn path, or the wall directly to the left. The black regions are the uncovered area when stitching frames from different time step.

tion module, relying on the model's capability to extract useful information. Meanwhile, the visual memory stitches the past frames together and integrates multi-modal inputs all into one single image, greatly improving the model and storage efficiency.

3.4 Hybrid Generation of Future Trajectories with Diffusion Model

The prediction part of the method heavily relies on a UNet diffusion model with multi-head self-attention layers between blocks. The self-attention layers help to relate the visual memory encoding to different parts of the trajectory, thus facilitating a deeper understanding of how humans interact with environmental features. The prediction is conditioned on a single 64-dimensional encoded visual memory, and full 100 steps of 6d pose trajectory in 20 Hz. We utilize the full 100 steps as we believe in some cases the ego-centric trajectory prediction task can violate pure Markovian assumption. For example, a different past trajectory could likely affect the visibility of the scene in visual memory, causing different exploration preferences in trajectory selection. We compared it to a pure Markovian assumption in the ablation.

The output of the diffusion model is a denoised sequence that contains both normalized 100 steps of ego-centric trajectory prediction and 100 steps of visual memory encoding. The visual memory encoding can be decoded to show how visual memory is expected to march forward by the model. Including the visual memory encoding in the diffusion process also provides a foundation for the model to base the predicted trajectory. In practice, we notice that the trajectory prediction samples are more stable with the visual memory prediction. In inference time, the model will take in the same condition and generate a batch of 15 different samples, forming a discrete distribution of the expected future ego trajectory modes.

We used Smooth L1 loss for diffusion model training. The VAE is trained on the same dataset as the diffusion model. For VAE, a combination of InfoLoss [25], Cross-entropy loss, and L1 loss is used. InfoLoss makes sure the encoding is highly correlated to the input, L1 loss is performed on RGBD channels, and Cross-entropy is performed on the one-hot semantic channels.

A conventional DDPM sampling method, while capable of producing high-quality predictions, operates at a pace that is impractical for applications requiring immediate responses, such as navigational aids or interactive systems. This limitation is further pronounced when attempting to generate a distribution of future trajectories, as multiple denoising sequences are necessary to produce a substantial number of samples, exacerbating the time constraints. Conversely, while DDIM offers a considerable acceleration in generating predictions, it does so at the expense of sample quality—a compromise that is untenable for applications where the fidelity of predicted trajectories directly impacts functionality and safety.

To address these challenges, we introduce a hybrid generation scheme that synergizes the strengths of both methods. Hybrid generation operates by initiating the generation process with a DDIM-like approach to quickly approximate the trajectory distribution, followed by a refinement phase using the DDPM framework. Essentially retaining the multimodal gradient landscape at the end of the diffusion process, ensuring that the final output maintains the intricate details and nuanced variations captured by a traditional DDPM without the accompanying latency.

4 Egocentric Navigation Dataset

Our Egocentric Navigation Dataset is collected within a university campus and its vicinity on human participants (under the approval of IRB-60675). The dataset comprises 34 carefully selected collections, each lasting approximately 7 minutes and spanning over 600 meters, designed to capture a wide range of interactions with the environment. These interactions are critical for testing and enhancing trajectory prediction models, particularly in densely populated or complex areas. The dataset is characterized by its diversity, encompassing various weather conditions (rain, sunny, overcast), surface textures (glass, solid, glossy, reflective, water), and environmental features (stairs, ramps, flat grounds, hills, off-road paths), alongside dynamic obstacles, including humans. Such variety ensures the dataset's applicability across a wide spectrum of egocentric navigation and prediction tasks.

Recorded at a 20Hz sampling rate, the dataset includes comprehensive state and visual information to capture the nuances in human behavior: 6 degrees of freedom (dof) torso pose in a global frame, leg joint angles (hips and knees of both legs), torso linear and angular velocity, and gait frequency. Visual data comprise aligned color and depth images, semantic segmentation masks, and visual memory frames generated to aid in trajectory prediction. This extensive collection amounts to 198 minutes of data (over 400GB). In practice, we find it

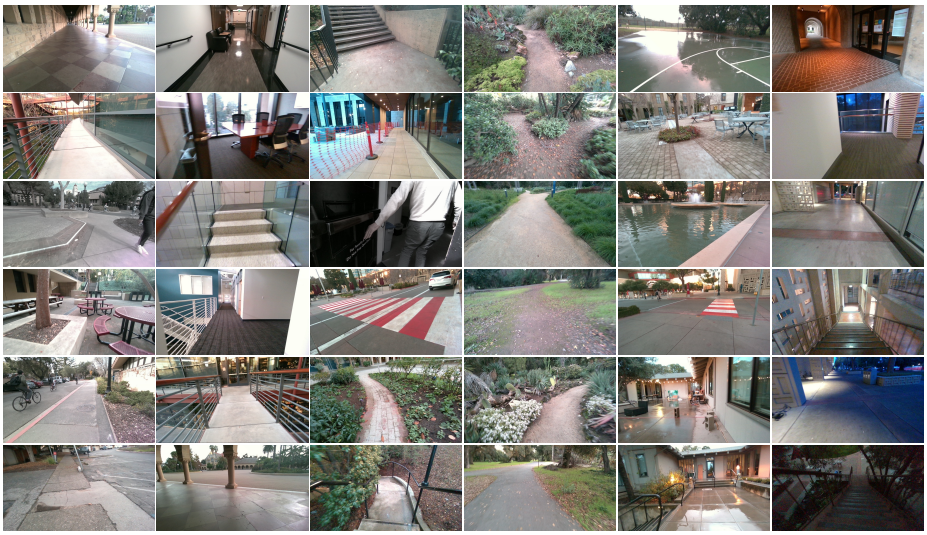

Fig. 5: The dataset contains a mix of weather, road, lighting, and traffic conditions

is possible to train a very high-quality model even with a smaller dataset size. Therefore we curated a high-quality pilot dataset with roughly 15% of the full data. This allows us to quickly iterate with faster training and an acceptable amount of performance degradation. The performance has been qualitatively compared in Sec. 5.

Acknowledging the limitations of existing datasets in this domain, such as the TISS dataset [11], our Egocentric Navigation Dataset addresses critical gaps by providing dense, high-frequency logs with rich visual and state information. Previous datasets have often suffered from sparse content, low logging frequencies, and lack of access to raw data or pre-trained decoder weights, hindering the development of sophisticated models and methods. To foster innovation and accessibility in egocentric trajectory prediction research, we are committed to open-sourcing our dataset following the de-identification of all faces within the data. We believe this approach will democratize access to high-quality data, enabling researchers to explore new methodologies and applications in the field.

5 Evaluation

5.1 Hardware Setup

We followed the hardware setup illustrated in this previous work [24]. To accommodate data collection in both indoor and outdoor environments, we opted for a setup that does not rely on GPS due to its indoor limitations and avoids IMU localization to minimize drift over extended periods. For precise localization, we employ an Intel Realsense T265 for SLAM-based tracking. We upgraded the stereo camera to an Intel Realsense D455, benefiting from its longer stereo

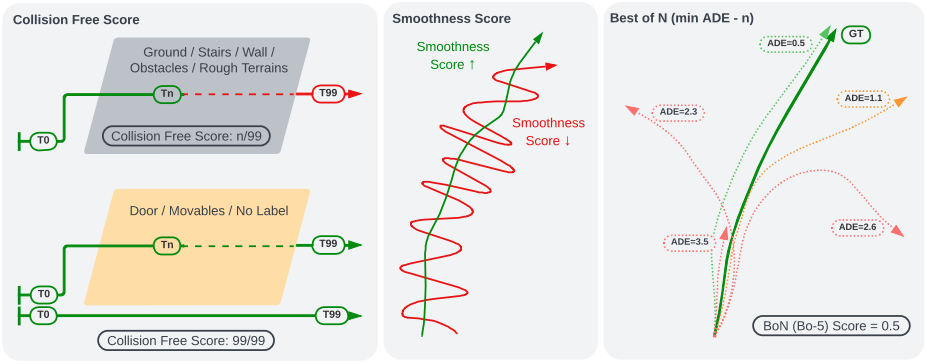

Fig. 6: Overview of metrics: The collision-free score (CFS) only considers points with selected semantics. All subsequent trajectories will be marked as collided. CFS and Smoothness are both higher the better, as opposed to Best of N.

baseline to enhance depth perception and extend the observable range. Considering the variability in lighting conditions, especially outdoors, we chose not to use LiDAR, despite its accuracy in optimal settings. The entire software stack, including real-time panorama generation and data storage, and physically powering the sensor is handled by an Apple Silicon MacBook Pro housed in a backpack. This setup ensures mobility and flexibility during data collection.

5.2 Evaluation Metrics

To comprehensively evaluate the predicted trajectories, we introduce three key metrics: collision-free score, smoothness, and best of N (including best of 1 as a subset for comparison with unimodal predictions). These metrics are designed to reflect the multifaceted nature of trajectory prediction, capturing aspects of physical feasibility, motion quality, and alignment with potential human preferences. An infographic summarizing these metrics is depicted in Fig. 6.

Collision-free Score This metric assesses the feasibility of the predicted trajectories by ensuring they do not intersect with impermeable objects within the environment. Given the absence of ground truth environmental meshes, we evaluate collisions against a discrete point cloud, defining a collision as the presence of more than 10 scene points within a 0.16m radius of the predicted position. The collision-free score tallies the number of consecutive prediction steps where the trajectory avoids collision with static obstacles such as the ground, stairs, walls, and rough terrains, excluding doors and movable objects. A higher score, up to the total number of prediction steps, indicates better performance.

Smoothness The smoothness metric quantifies the continuity and fluidity of movement in the generated trajectories. Calculated as the reciprocal of the sum

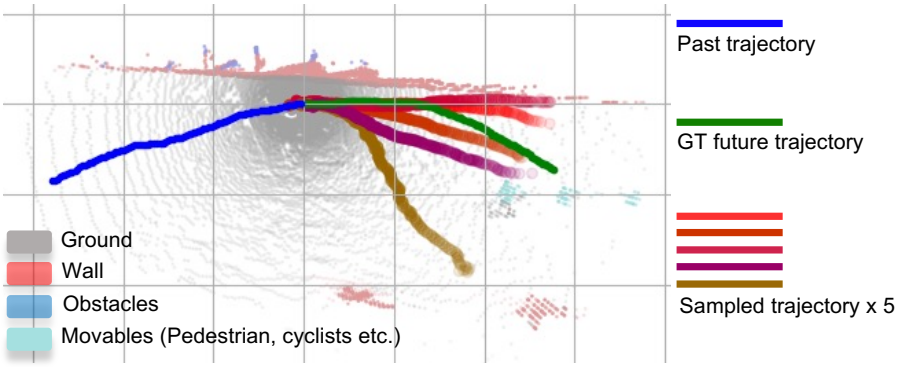

Fig. 7: BEV: Typical input and output are visualized top-down. Five trajectories are predicted in this case, each with a different shade of red. Visual memory is projected to a point cloud and color-coded same as semantic.

of the mean absolute errors (MAE) of speed and acceleration compared to ground truth, this metric relates higher values to smoother trajectories. This adjustment ensures that higher scores correspond to desirable attributes, aligning with the intuitive understanding that smoother trajectories are preferable.

$$Smoothness = 1/(\frac{\sum_{i=1}^{n}|V_i - \hat{V}_i|}{n} + \frac{\sum_{i=1}^{n}|a_i - \hat{a}_i|}{n}) \tag{2}$$

Best of N & Best of 1 Recognizing the inherent multiplicity of valid trajectories for any given scenario, the Best of N metric evaluates the model's ability to encapsulate the ground truth trajectory within its distribution of predictions. This approach mirrors the minADE-K metric used in autonomous vehicle research, focusing on the minimum average displacement error across a set of K predictions. The Best of 1 metric complements this by assessing the accuracy of a single prediction, facilitating comparison with models that do not generate a multimodal distribution. It also shows how well the estimated distribution aligns with the ground truth under the assumption that the ground truth mode should have the highest likelihood in the given scenario. Together, these metrics ensure a balanced evaluation of the model's predictive capacity and its versatility in capturing the range of plausible ego-motions.

5.3 Prediction Evaluation

Our evaluation focuses on assessing the model's capacity for future trajectory prediction and its ability to generate expected future visual memory encodings, as detailed in Sec. 3. Operating over a 100-step (5-second) prediction horizon—aligned with the typical visibility range (8-meter radius) and walking speeds—our method ensures the generation of predictions that are both relevant

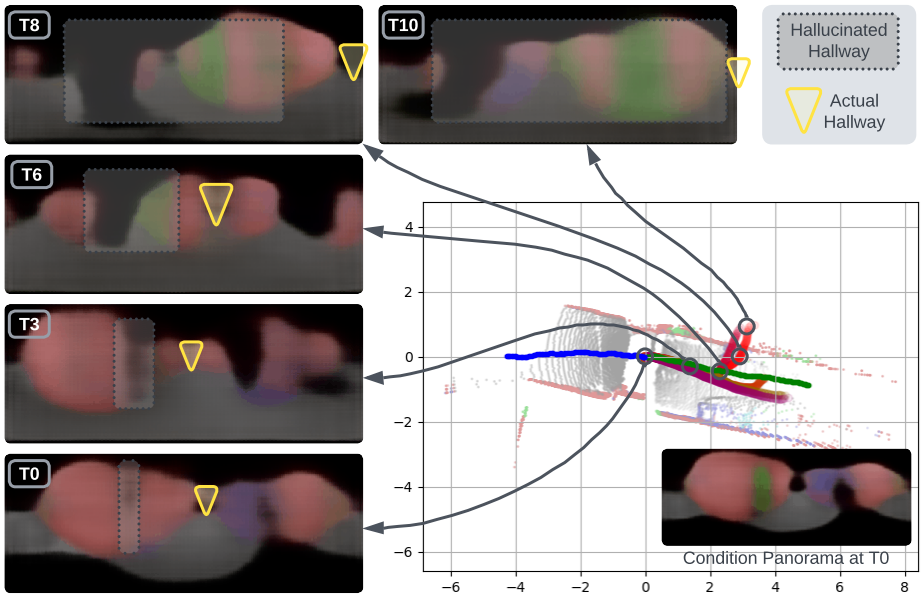

Fig. 8: Analysis of fail prediction: A failed prediction (bright red) is picked to analyze the failure mode. Predicted visual memory frames are sampled along the trajectory to show the model's expectation. Visual memory is always centered to forward direction, turning can be observed by following the centerline. BEV plot: Blue trajectory is past, and green is the ground truth. Point clouds follow semantic color codes.

and plausible within the observable environment. The inclusion of future visual memory alongside trajectory predictions enhances our ability to visually validate the model's anticipations of scene evolution, providing a robust cross reference for trajectory prediction and insight into the model's typical failure modes.

A pertinent example of this evaluation is depicted in Fig. 8, where visual memories associated with the red trajectory prediction reveal a challenge commonly encountered by generative models: the occasional "hallucination" of nonexistent features in scenarios of occluded vision or gaps in the observable scene. For instance, as a person navigates a hallway, gaps in conditioning visual memories might occur due to changes in view angle or partial occlusions. At a specific timestep (T3), the model inaccurately anticipates that approaching a gap would reveal a larger, unexplored space (as in T6), leading it to "hallucinate" an extension of the hallway based on patterns observed in the training data. This results in the model predicting a trajectory that erroneously veers into this fictitious extension (T8), as evidenced by the gradual shift of the actual hallway from the center of the predicted path.

5.4 Ablations

To test the effectiveness of individual modules proposed in our method, we conducted the following ablation studies. The individual ablations and correspond-

ing metrics are summarized in Tab. 1. As mentioned in Sec. 4, we have a full dataset for the best possible performance and a curated pilot dataset for quick iteration. All of the ablation training, unless specifically mentioned, are trained on the pilot dataset. The entries in the table have been sorted by the collision-free score in the second column. The ground truth trajectory did not receive a perfect score (99) in collision avoidance, due to the inaccuracies in the SLAM pose estimation and depth artifacts from the stereo camera. In rare occasions, ground truth trajectory collides with a part of the point cloud that appears a considerable distance away from the ground truth surface of the scene.

Scalability To quantify the difference between training on a large dataset and the pilot dataset, we trained two versions of the baseline one on each dataset. The full dataset model is better on all of the evaluation metrics as expected. However the gap is quite small on the collision metric, we believe the collision avoidance performance is highly dependent on the accuracy of the point cloud, otherwise the model would think that the ground truth demonstration has penetrated some points despite the points being from artifacts and misestimations. On the other hand, the added samples have greatly improved the variance of the generation and hence better Best of n scores in general.

Transformer UNet vs UNet Adding in the multi-head self-attention layers (MHSA) between down and up blocks in UNet significantly blows up the size of the model, fortunately, the inference and training speeds have remained largely unchanged. Without attention layers, the most significant degradation in performance is in collision avoidance. This corroborates our expectation that the MHSA layers help to relate encoded visual memory with trajectory samples.

Hybrid Generation As proposed in Sec. 3, hybrid generation can greatly speed up generation at the expense of a slight decrease in quality (as quantified in ablation), up to 50 trajectory predictions per second on a single Nvidia A5000 GPU. In this experiment, our hybrid generation uses 20 steps of DDIM, followed by 10 steps of DDPM to refine the generation. Compared to a full 1000-step DDPM baseline, this method is 33 times faster, but the quantitative metrics only decreased slightly. To show the effectiveness, we also include the same 30-step generation result except from DDIM. Both collision and smoothness scores were negatively impacted especially the smoothness score. The best-of-n metrics remain similar, showing that while the DDIM is converging in the right direction, it does not converge to a strong solution like DDPM.

Past Condition Length To achieve the best possible performance, we are using a non-Markovian assumption when performing prediction. This requires us to provide a long history of past trajectories. We are curious as to how well the Markovian assumption will hold for this task and are interested in potentially reducing the number of inputs to the model for more efficient use of computation. Therefore we trained a variant that only takes in 3 steps of past trajectory steps, essentially only providing information up to the order of acceleration. The resulting model did not suffer too much from the change, showing that the Markovian assumption largely holds for most of the scenarios.

Table 1: Metric comparison table across ablations

Ablation	| Collision ↑|	Smoothness ↑|	Best of 1 ↓|	Best of 15↓
Ground Truth	97.6	-	-	-
Ours (Full dataset)	**90.6**	**4.76**	**0.81**	**0.41**
w/o visual memory prediction	*89.3*	3.13	0.91	0.50
Ours (Pilot Dataset)	89.2	2.04	*0.87*	*0.47*
w/ Markovian past state	88.8	1.56	1.04	0.52
w/ Hybrid generation (I20 P10)	88.7	2.17	0.89	0.49
w/o attention	86.6	2.78	1.00	0.49
w/ DDIM generation (n=30)	85.3	0.46	0.92	0.52
w/o semantic (RGBD only)	84.1	*4.17*	0.91	0.53
w/o visual input (Traj only)	82.5	2.04	1.19	0.48

Semantic Segmentation To test the effectiveness of semantic segmentation. The VAE and diffusion model are trained on the same dataset except without a semantic channel. We see a significant drop in collision avoidance performance. This stems from the same origin as the quality issue we mentioned in Scalability above. Without semantics, the model will not be able to differentiate between penetrating the door and the wall and, thus tends to ignore the geometric constraints provided in the visual memory.

Visual Input Empirically, the removal of the visual input results in the worst-performing model. This emphasizes again the effectiveness of contextual structure provided by the visual memory. Upon inspecting the results, many of the generated samples are simply memorizing what was in the training set and the distribution is incorrect (which also corresponds to a high best-of-1 score as mentioned in Evaluation Metrics of Sec. 5).

5.5 Comparison with Baselines

The proposed method is compared to existing methods LSTM and CXA Transformer. For consistency, all three methods are based on the same pre-trained VAE. And they are all trained and evaluated on the pilot dataset. Results are shown in Tab. 2. We would like to highlight that some key differences between our model and the baselines are: 1) Our model does not rely on auto-regressive generation and, therefore less affected by compounding error problems. Our model also generally requires less pass through the generation model, which means faster speed. 2) Our model is inherently based on a random process, therefore it learns an implicit distribution of preferences instead of simply an average.

VAE-LSTM One of the most common paradigms to the trajectory prediction task in the existing literature [11,24] is to use an encoder head to understand the visual input, and then pass the state information along with encoded visual cues into a recurrent model to auto-regressively predict the future states. The quality of the output is usually pretty good. The biggest drawback of this approach is that the prediction is deterministic, as the variance only exists in the

Table 2: Metric comparison table across baselines

Model	Collision ↑	Smoothness ↑	Best of 1 ↓	Best of 15 ↓
CXA Transformer	80.3	3.70	1.25	N/A
LSTM-VAE	84.5	**0.09**	1.01	N/A
Ours	**89.2**	2.04	**0.87**	**0.472**

encoding process of the visual input, not in the trajectory latent space. And since there is only one deterministic prediction in each scenario, there is zero guarantee of the temporal consistency of the prediction between consecutive scenarios. Despite the drawbacks, the auto-regressive model excels at generating smooth motion as every step is based on the previous predictions, which is also confirmed by the smoothness score. We implemented the LSTM baseline according to the receipt provided in this egocentric trajectory prediction paper [24]. The LSTM model in the baseline performs surprisingly well at avoiding obstacles, given its tiny size of fewer than 1 million parameters.

CXA-Transformer One of the newer approaches is to use the transformer instead of the recurrent modules [11]. The transformer models, at its core, rely on attention mechanisms to process the sequence so it suffers less from forgetting and improves computational efficiency for scaling. One of the most recent works aiming to tackle egocentric trajectory prediction in a similar setup proposes a Cascaded Cross-Attention(CXA) Transformer. It uses a transformer encoder for each modality of input including surrounding people's pose, semantic segmentation, and past trajectories, and a cross-attention module to fuse in multi-modal information. Similar to other transformer models, it uses a transformer decoder to produce a prediction of the most likely future path auto-regressively. CXA transformer is implemented strictly as outlined and uses hyper-parameters in the paper. In training and evaluation, we noticed a strong tendency to overfit to the training set, despite adding a significant amount of dropouts and many hyper-parameter tuning. The best result we can get is shown in the table.

6 Conclusion

This paper presented a pipeline to predict future trajectories using diffusion model conditioned on the past trajectory and scene semantics. We show that it benefits from semantic segmentation, and outperforms existing methods. Limitations of our method include the current assumption that the scene is static, despite it is implicitly handled by semantic segmentation, it can still get quite challenging when there are numerous people around.

Future work in this area includes employing monocular depth estimation models for dense depth and more informative scene capture. Accounting for dynamic obstacles in the scene. Upgrading the latent diffusion model to consistency model [19] for real-time generation is also desirable especially when deploying the model to mobile platforms.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: Human Trajectory Prediction in Crowded Spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–971. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.110, http://ieeexplore.ieee.org/document/7780479/ 3
2. Bertasius, G., Chan, A., Shi, J.: Egocentric Basketball Motion Planning from a Single First-Person Image (Mar 2018), http://arxiv.org/abs/1803.01413, arXiv:1803.01413 [cs] 3
3. Chai, Y., Sapp, B., Bansal, M., Anguelov, D.: MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction (Oct 2019), http://arxiv.org/abs/1910.05449, arXiv:1910.05449 [cs, stat] 3
4. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks (Mar 2018), http://arxiv.org/abs/1803.10892, arXiv:1803.10892 [cs] 3
5. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models (Dec 2020). https://doi.org/10.48550/arXiv.2006.11239, http://arxiv.org/abs/2006.11239, arXiv:2006.11239 [cs, stat] 3
6. Karkus, P., Ivanovic, B., Mannor, S., Pavone, M.: DiffStack: A Differentiable and Modular Control Stack for Autonomous Vehicles (Dec 2022). https://doi.org/10.48550/arXiv.2212.06437, http://arxiv.org/abs/2212.06437, arXiv:2212.06437 [cs] 3
7. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by Example. Computer Graphics Forum **26**(3), 655–664 (2007). https://doi.org/10.1111/j.1467-8659.2007.01089.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2007.01089.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2007.01089.x 3
8. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using Denoising Diffusion Probabilistic Models (Aug 2022). https://doi.org/10.48550/arXiv.2201.09865, http://arxiv.org/abs/2201.09865, arXiv:2201.09865 [cs] 4
9. Mercat, J., Gilles, T., Zoghby, N.E., Sandou, G., Beauvois, D., Gil, G.P.: Multi-Head Attention for Multi-Modal Joint Vehicle Motion Forecasting (Dec 2019), http://arxiv.org/abs/1910.03650, arXiv:1910.03650 [cs] 3
10. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Feb 2024). https://doi.org/10.48550/arXiv.2304.07193, http://arxiv.org/abs/2304.07193, arXiv:2304.07193 [cs] 5
11. Qiu, J., Chen, L., Gu, X., Lo, F.P.W., Tsai, Y.Y., Sun, J., Liu, J., Lo, B.: Egocentric Human Trajectory Forecasting With a Wearable Camera and Multi-Modal Fusion. IEEE Robotics and Automation Letters **7**(4), 8799–8806 (Oct 2022). https://doi.org/10.1109/LRA.2022.3188101, https://ieeexplore.ieee.org/document/9813561/ 8, 13, 14
12. Rempe, D., Luo, Z., Peng, X.B., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., Litany, O.: Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion (Apr 2023). https://doi.org/10.48550/arXiv.2304.01893, http://arxiv.org/abs/2304.01893, arXiv:2304.01893 [cs] 3

13. Rempe, D., Philion, J., Guibas, L.J., Fidler, S., Litany, O.: Generating Useful Accident-Prone Driving Scenarios via a Learned Traffic Prior. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17284–17294. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01679, https://ieeexplore.ieee.org/document/9880074/ 3
14. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 549–565. Lecture Notes in Computer Science, Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_33 3
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models (Apr 2022). https://doi.org/10.48550/arXiv.2112.10752, http://arxiv.org/abs/2112.10752, arXiv:2112.10752 [cs] 4
16. Schreiber, M., Hoermann, S., Dietmayer, K.: Long-Term Occupancy Grid Prediction Using Recurrent Neural Networks (Jun 2019), http://arxiv.org/abs/1809.03782, arXiv:1809.03782 [cs] 3
17. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics (Nov 2015). https://doi.org/10.48550/arXiv.1503.03585, http://arxiv.org/abs/1503.03585, arXiv:1503.03585 [cond-mat, q-bio, stat] 3
18. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models (Oct 2022). https://doi.org/10.48550/arXiv.2010.02502, http://arxiv.org/abs/2010.02502, arXiv:2010.02502 [cs] 4
19. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency Models (May 2023). https://doi.org/10.48550/arXiv.2303.01469, http://arxiv.org/abs/2303.01469, arXiv:2303.01469 [cs, stat] 14
20. Strohbeck, J., Belagiannis, V., Muller, J., Schreiber, M., Herrmann, M., Wolf, D., Buchholz, M.: Multiple Trajectory Prediction with Deep Temporal and Spatial Convolutional Neural Networks. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1992–1998. IEEE, Las Vegas, NV, USA (Oct 2020). https://doi.org/10.1109/IROS45743.2020.9341327, https://ieeexplore.ieee.org/document/9341327/ 3
21. Su, S., Hong, J.P., Shi, J., Park, H.S.: Predicting Behaviors of Basketball Players from First Person Videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1206–1215. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.133, http://ieeexplore.ieee.org/document/8099616/ 3
22. Tang, Y.C., Salakhutdinov, R.: Multiple Futures Prediction (Dec 2019), http://arxiv.org/abs/1911.00997, arXiv:1911.00997 [cs, stat] 3
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (Aug 2023). https://doi.org/10.48550/arXiv.1706.03762, http://arxiv.org/abs/1706.03762, arXiv:1706.03762 [cs] 3
24. Wang, W., Raitor, M., Collins, S., Liu, C.K., Kennedy III, M.: Trajectory and Sway Prediction Towards Fall Prevention (Mar 2023). https://doi.org/10.48550/arXiv.2209.11886, http://arxiv.org/abs/2209.11886, arXiv:2209.11886 [cs] 8, 13, 14
25. Zhao, S., Song, J., Ermon, S.: InfoVAE: Information Maximizing Variational Autoencoders (May 2018). https://doi.org/10.48550/arXiv.1706.02262, http://arxiv.org/abs/1706.02262, arXiv:1706.02262 [cs, stat] 7

7 Appendix

7.1 Dataset Detail

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

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

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

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

7.2 Diffusion Model Detail

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

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

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

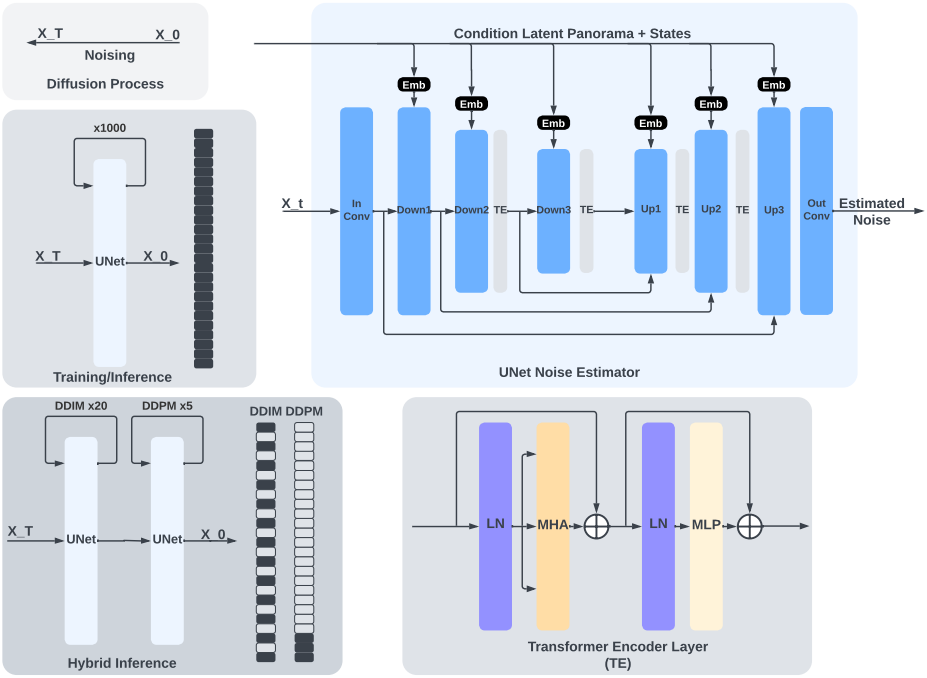

Fig. 9: Diffusion Model: Architecture and hybrid generation details

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

Where:

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

7.3 Collision-Free Metric Detail

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