Enhanced Environmental Context Encoding for Accurate Trajectory Prediction in Intralogistics

Alexander Prutsch
Institute of Visual Computing
Graz University of Technology
Graz, Austria
alexander.prutsch@tugraz.at

Horst Possegger
CD Laboratory for Embedded Machine Learning,
Institute of Visual Computing
Graz University of Technology
possegger@tugraz.at

Abstract

Trajectory prediction is an essential component of the perception stack in autonomous mobile robots (AMRs). AMRs operate in complex environments where their movements are influenced by various environmental elements, such as racks and storage locations. Therefore, accurate and efficient trajectory prediction for intralogistics requires detailed environment modeling that goes beyond the lane-based context commonly used for road traffic. We propose a new environment context encoder that can be seamlessly integrated into state-of-the-art motion forecasting models. Our approach, tailored to the specific challenges of intralogistics, achieves highly accurate predictions using efficient baseline networks.

1 Introduction

Autonomous mobile robots (AMRs) play a major role in modern intralogistics as they are commonly used to transport cargo in complex environments like warehouses and production facilities. AMRs require accurate trajectory prediction to navigate crowded intralogistics environments efficiently and safely. Anticipating the movements of other traffic participants allows AMRs to proactively plan maneuvers, which prevents operational delays from costly deadlocks and protects warehouse workers.

While trajectory prediction is well-studied for road traffic, *e.g.,* [1, 2, 3, 4, 5, 6, 7], intralogistics presents distinct challenges. Warehouse vehicles can execute more diverse and complex maneuvers due to their wheel geometry [8]. Additionally, lane graphs, which are commonly used to model map information in road traffic, provide only weak guidance in the intralogistics domain, as vehicles follow driving lanes less strictly within warehouses and production facilities. Furthermore, environmental elements heavily influence driving behavior; *e.g.,* a vehicle entering a rack aisle has a high probability of abruptly changing speed to initiate load handling. Consequently, explicitly modeling these map elements is critical for robust trajectory prediction in intralogistics.

Due to different environment types, vehicle characteristics, and onboard resources, trajectory prediction models designed for autonomous driving cannot be directly transferred to the AMR domain. We propose a new environment encoder module to enable accurate trajectory prediction in intralogistics. Our **e**nvironment **c**on**t**ex**t** (ECTX) encoder processes information on diverse map elements like rack positions, charging stations and gates, enabling highly accurate predictions in scenarios where lanes only offer a weak prior. By directly utilizing widely available semantic warehouse maps, ECTX bypasses the need for computationally heavy LiDAR data [6, 5] and requires no application-specific fine-tuning using LiDAR data. Integrating our ECTX module into two strong, compact transformer-based baselines, Forecast-MAE [9] and EMP [10], yields improved results for trajectory prediction in complex intralogistics driving scenarios. The additional environmental information significantly improves the trajectory prediction accuracy while adding only little overhead to the networks. We demonstrate the effectiveness of our ECTX by evaluating it on a custom large-scale dataset for motion

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

prediction of different intralogistics vehicles, *e.g., reach trucks* and *order pickers*. Our approach demonstrates strong performance on long-term prediction horizons, giving AMRs sufficient time to react to the prediction and to use the output to perform smooth, proactive driving maneuvers.

2 Trajectory Prediction With Enhanced Environment Encoding

Modern trajectory prediction typically relies on separate agent and lane encoders prior to scene encoding. To capture the unique dynamics of intralogistics, we introduce a plug-and-play third module (ECTX), which integrates critical environmental context into baseline architectures, *i.e.,* [9, 10].

Map Modeling: We extract lane segments and environmental elements from standard warehouse floor plans. Lane segments are sampled and split into similar-sized, fixed-point chunks. Crucially, we extract key environmental elements that dictate driving behavior, *e.g.,* racks (indicating potential load handling), non-driveable areas (walls, machinery), and free areas (where vehicles may cut corners).

Agent and Lane Encoding: Initially, agent history and lane shapes are encoded individually using local coordinate systems. For agent motion, baselines output a feature vector for each agent using either neighborhood attention [11] (Forecast-MAE [9]) or standard self-attention (EMP [10]). Lane geometries are processed in both baselines using PointNet-like architectures [12].

Environment Encoding (ECTX): We encode points sampled from the environment polygons using a small PointNet-like network [12]. Unlike the lane encoder, which extracts line features, our environment encoder learns the shapes of polygons. It outputs an environment matrix where each row represents a distinct map area.

Scene Encoding and Trajectory Decoding: As both baselines [9, 10] use token-type independent self-attention for scene encoding, we can seamlessly concatenate the environment tokens with the agent and lane tokens to form a unified scene context. To preserve spatial and categorical relationships, we add global positional embeddings [9] and type embeddings for both vehicle and map classes. Finally, a set of trajectory hypotheses is decoded using the baselines respective architecture (an MLP for Forecast-MAE and EMP-M, or a DETR-like [13] decoder for EMP-D).

3 Experimental Setup

Dataset: We conduct our evaluations on a large-scale dataset generated using NVIDIA Omniverse™. It is recorded in virtual warehouse environments, which are designed based on real-world layouts and traffic situations. We use CAD models of different real-world intralogistics vehicles and apply custom vehicle controllers to obtain highly realistic motion patterns. The maps in our dataset contain lane topology implemented as directed graphs and detailed information on intralogistics environment elements. These include charging stations, free areas, rack locations, gates, non-driveable areas (static obstacles), and storage locations. Overall, our dataset features 267,146 total scenarios across two virtual environments: one for training (94,621) and validation (63,605), and a second environment for testing (108,920). The intentionally limited training set reflects real-world data scarcity.

Each scenario spans 11 seconds (5 s history, 6 s prediction horizon), yielding a challenging median future trajectory length of 10.45 m (max 13.60 m). The dataset includes various vehicle types, *e.g.,* reach trucks, forklifts, and order pickers, to capture the different driving dynamics of each type. Compared to autonomous driving data, our dataset includes difficult intralogistics-specific movements like load handling, on-the-spot rotations, and reversing. For each scenario, we sample all neighboring map elements and agents within a radius of 25 m as model input, which fully captures the relevant context for intralogistics driving speeds.

Implementation Details: For both baselines [9, 10], we evaluate the original architectures (latent feature dimension $D = 128$) alongside a smaller version with $D = 64$ and shallower encoders. The baseline models are designed for autonomous driving applications, where large-scale datasets are available. This model size reduction mitigates potential overfitting on the smaller datasets typical of intralogistics and facilitates deployment on embedded hardware. Models are trained on a single NVIDIA V100 GPU for 60 epochs with a batch size of 128. We do not use data augmentation and optimize using AdamW [14] with gradient clipping and weight decay. The learning rate undergoes a 10-epoch linear warm-up (1×10^{-6} to 8×10^{-5}) before a cosine decay schedule back to 1×10^{-6}. Following our baselines [9, 10], the training objective combines a regression loss, a classification loss

Table 1: Results on our intralogistics dataset grouped by baseline model and sorted in descending order by test **brier-minFDE$_4$**. Models marked with * denote reduced architecture configurations, which are also more suitable for AMR deployment. Displacement errors reported in meters.

Method	Test Set MR$_4$	minADE$_1$	minFDE$_1$	minADE$_4$	minFDE$_4$	**brier-minFDE$_4$**
EMP-M [10]	0.156	1.12	2.83	0.57	1.21	1.56
EMP-M*	0.182	1.09	2.69	0.57	1.22	1.52
EMP-M*+ECTX	0.155	1.08	**2.67**	0.55	1.17	1.46
EMP-D [10]	0.140	1.12	2.84	0.51	1.07	1.41
EMP-D*	0.129	1.11	2.76	0.51	1.06	1.35
EMP-D*+ECTX	0.129	1.15	2.83	0.50	1.03	1.33
Forecast-MAE [9]	**0.117**	1.16	2.97	0.48	**0.98**	1.34
Forecast-MAE*	0.126	**1.06**	**2.67**	0.49	1.03	1.32
Forecast-MAE*+ECTX	**0.117**	**1.06**	2.70	**0.47**	**0.98**	**1.27**

Table 2: Ablation study on the influence of different encoder module using Forecast-MAE*. The experiment marked with ✓† uses only a single encoder for environment and lane data.

Context Encoder Lanes	ECTX	Test Set MR$_4$	minADE$_4$	minFDE$_4$	**brier-minFDE$_4$**
✓†		0.323	0.84	1.72	2.03
✗	✗	0.251	0.67	1.49	1.83
✗	✓	0.188	0.64	1.38	1.70
✓	✗	0.126	0.49	1.03	1.32
✓	✓	**0.117**	**0.47**	**0.98**	**1.27**

to score the multiple trajectory hypotheses, and an auxiliary loss that predicts a single future for all non-focal agents in the scene.

4 Results and Conclusions

We present detailed evaluations on our custom intralogistics dataset by comparing three baseline models with and without our ECTX encoder. We evaluate our models using standard trajectory prediction metrics [15, 16, 17]: minADE$_K$, minFDE$_K$, brier-minFDE$_K$, and MR$_K$. To capture diverse future movements while maintaining a compact representation suitable for AMR control systems, our models output 4 trajectory hypotheses. We compute these metrics for both the most probable prediction ($K = 1$) and the full set of predictions ($K = 4$).

Evaluation on Intralogistics Dataset: Table 1 compares our baselines, Forecast-MAE [9] and EMP-M/D [10], with and without our new ECTX module on our custom intralogistics dataset. For all three models, the integration of ECTX significantly improves trajectory prediction accuracy on our test set, leading to highly accurate overall results. Furthermore, our reduced-capacity configurations (marked by *) outperform the original architectures, better matching the limited dataset size while suiting the computational constraints of embedded AMR hardware. Consistent with the results from [10] on AV2 [17], EMP-D outperforms EMP-M due to its more sophisticated decoder architecture. The intralogistics scenarios feature significantly fewer agents per scene, where the neighborhood attention-based [11] agent encoder from Forecast-MAE brings an advantage over the pure transformer version from EMP. Furthermore, comparing the $K = 1$ and $K = 4$ metrics reveals that ECTX substantially improves multiple-hypothesis generation, demonstrating that explicit environmental context is crucial for accurately modeling multi-modal behavior.

Figure 1 compares the predictions of EMP-D* without and with ECTX on two example scenarios from our test dataset. In the first scenario (top row), a vehicle navigates a *rack pre-zone*, where it can either enter a rack aisle or continue in the open area. Without the environmental context of the rack positions, the baseline EMP-D* predicts an invalid left turn that would lead to a collision with a rack. Adding the rack positions using our ECTX encoder leads to well-suited trajectory predictions. In the second scenario (bottom row), a vehicle is leaving an aisle having the option to either go left or right. Using only lane data as context, the baseline model predicts a corner-cutting maneuver that would

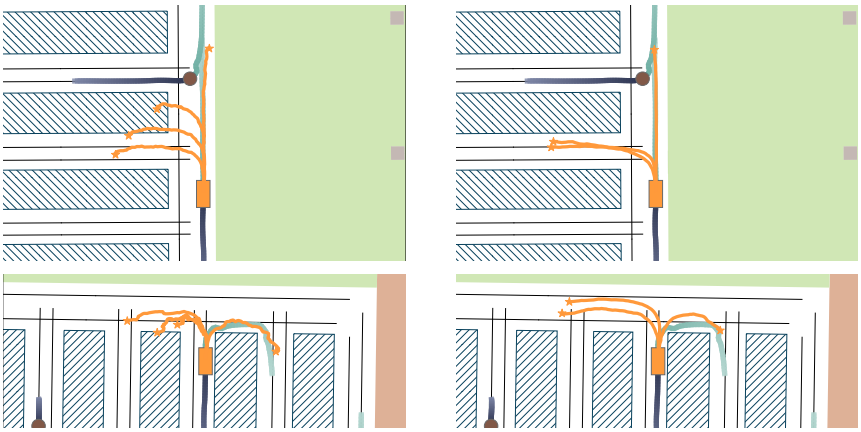

Figure 1: Both image pairs show scenarios from our custom intralogistics dataset. The left images show **predictions** from our baseline EMP-D* model and the right images from EMP-D*+ECTX. In all images also the **ground truth trajectory** is shown. Using ECTX the possible turn inside the rack aisles much better aligns with the given environment structures. For the second example using ECTX the predicted trajectories do not interfere with the **racks** as the corners are not cut.

Table 3: Latency comparison for deployment on robotic hardware, alongside measurements from a reference desktop GPU. We report inference latency for predicting all agents in the most complex test scene (8 vehicles, 110 context elements).

Method	NVIDIA Jetson		NVIDIA	Model
	Orin Nano	AGX Orin	V100	Parameters
EMP-D*+ECTX	38 ms	15 ms	28 ms	1.2M
EMP-M*+ECTX	37 ms	15 ms	22 ms	791K

collide with a static obstacle. Once again, incorporating our environment encoder resolves this issue, ensuring the predicted trajectories remain physically viable and collision-free.

Ablation Study: Table 2 details an ablation study evaluating the influence of lane and environment context using Forecast-MAE [9]. As expected, map-free prediction (agent history only) performs worst, as it cannot anticipate structural maneuvers like turns and predictions are limited to different driving motion patterns. Using only the environment context from ECTX as input, trajectory prediction accuracy significantly improves. This confirms that using the map elements provides a valuable input for trajectory prediction. As expected, the addition of lane data yields the best results overall. We also conduct an experiment where we utilize a single encoder module to encode both lane polylines and environment polygons simultaneously. Processing both lane polylines and environment polygons through a single encoder degrades performance, as the model fails to learn proper context guidance. This confirms that distinct spatial modalities require dedicated encoder modules.

Resource Analysis: We evaluate the real-world inference latency of our EMP-based models for predicting all agents in the most complex scene in our test set. For hardware, we use two types of NVIDIA Jetson devices, which are designed for robotics applications, and include a comparison with a standard NVIDIA GPU. The results in Table 3 highlight that our approach is very well suited for real-time processing on AMR hardware. Forecast-MAE could not be tested in this evaluation setting, because it uses neighborhood attention blocks [11], which are not supported for export to ONNX.

Conclusions: To solve trajectory prediction in complex intralogistics environments, we introduce the **e**nvironment **c**on**t**e**x**t (ECTX) encoder, a versatile extension for state-of-the-art trajectory prediction models. The ECTX encoder captures detailed map information on intralogistics specific elements like racks. Extensive evaluations demonstrate that integrating ECTX significantly enhances baseline accuracy. Furthermore, our findings emphasize that for custom robotic domains, specialized compact architectures consistently outperform standard, large-scale models designed for autonomous driving.

Acknowledgments: We gratefully acknowledge the financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association.

References

[1] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal Motion Prediction with Stacked Transformers. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7577–7586, 2021.

[2] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion Transformer with Global Intention Localization and Local Movement Refinement. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-Centric Trajectory Prediction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] Xishun Wang, Tong Su, Fang Da, and Xiaodong Yang. ProphNet: Efficient Agent-Centric Motion Forecasting with Anchor-Informed Proposals. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] Yiqian Gan, Hao Xiao, Yizhe Zhao, Ethan Zhang, Zhe Huang, Xin Ye, and Lingting Ge. MGTR: Multi-Granular Transformer for Motion Prediction with LiDAR. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2024.

[6] Kan Chen, Runzhou Ge, Hang Qiu, Rami Ai-Rfou, Charles R Qi, Xuanyu Zhou, Zoey Yang, Scott Ettinger, Pei Sun, Zhaoqi Leng, et al. WOMD-LiDAR: Raw Sensor Dataset Benchmark for Motion Forecasting. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2024.

[7] Yang Zhou, Hao Shao, Letian Wang, Steven L. Waslander, Hongsheng Li, and Yu Liu. SmartRefine: A Scenario-Adaptive Refinement Framework for Efficient Motion Prediction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] Alexander Prutsch, Horst Possegger, and Horst Bischof. Action-By-Detection: Efficient Forklift Action Detection for Autonomous Mobile Robots in Warehouses. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2024.

[9] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-MAE: Self-supervised Pre-training for Motion Forecasting with Masked Autoencoders. In *Proc. of the IEEE/CVF Conference on Computer Vision (ICCV)*, 2023.

[10] Alexander Prutsch, Horst Bischof, and Horst Possegger. Efficient Motion Prediction: A Lightweight & Accurate Trajectory Prediction Model With Fast Training and Inference Speed. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[11] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood Attention Transformer. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[14] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

[15] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3D Tracking and Forecasting with Rich Maps. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting. In *Proc. of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.