Learning Causally-Aware Representations of Multi-Agent Interactions

Yuejiang Liu* **Ahmad Rahimi*** **Po-Chien Luan*** **Frano Rajič** **Alexandre Alahi**
École Polytechnique Fédérale de Lausanne (EPFL)
{firstname.lastname}@epfl.ch

Abstract

Modeling spatial-temporal interactions between neighboring agents is at the heart of multi-agent problems such as motion forecasting and crowd navigation. Despite notable progress, it remains unclear to which extent modern representations can capture the causal relationships behind agent interactions. In this work, we take an in-depth look at the causal awareness of the learned representations, from computational formalism to controlled simulations to real-world practice. First, we cast doubt on the notion of non-causal robustness studied in the recent CausalAgents benchmark [54]. We show that recent representations are already partially resilient to perturbations of non-causal agents, and yet modeling indirect causal effects involving mediator agents remains challenging. Further, we introduce a simple but effective regularization approach leveraging causal annotations of varying granularity. Through controlled experiments, we find that incorporating finer-grained causal annotations not only leads to higher degrees of causal awareness but also yields stronger out-of-distribution robustness. Finally, we extend our method to a sim-to-real causal transfer framework by means of cross-domain multi-task learning, which boosts generalization in practical settings even without real-world annotations. We hope our work provides more clarity to the challenges and opportunities of learning causally-aware representations in the multi-agent context while making a first step towards a practical solution.

1 Introduction

Modeling multi-agent interactions with deep neural networks has made great strides in the past few years [2, 12, 21, 22, 29, 44, 58, 65, 66]. Yet, existing representations still face tremendous challenges in handling changing environments: they often suffer from substantial accuracy drops under mild environmental changes [4, 37] and require a large number of examples for adaptation to new contexts [31, 49]. These challenges are arguably rooted in the nature of the learning approach that seeks statistical correlations in the training data, regardless of their stability and reusability across distributions [4, 7]. One promising solution is to build more *causally-aware representations* – latent representations that are capable of capturing the invariant causal dependencies behind agent interactions.

However, discovering causal knowledge from observational data is often exceptionally difficult [60, 61]. Most prior works resort to additional information, such as structural knowledge [10, 43] and domain labels [4, 26, 38]. While these attempts have been shown effective in certain out-of-distribution scenarios, they still fall short of explicitly accounting for causal relationships between interactive agents. More recently, CausalAgents [54] made an effort to collect annotations of agent relations (causal or non-causal) in the Waymo dataset [18], providing a new benchmark focused on the robustness issue under non-causal agent perturbations. Nevertheless, the reason behind the robustness issue remains unclear, as does the potential use of the collected annotations for representation learning.

*equal contribution

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Figure 1: Illustration of multi-agent interactions. The behavior of the ego agent is causally influenced by some neighbors. We study the challenges and potential for modeling the causal relations between interactive agents.

The goal of this work is to provide an in-depth analysis of the challenges and opportunities of learning causally-aware representations in the multi-agent context. To this end, we first take a critical look at the recent CausalAgents [54] benchmark. We find that its labeling procedure and evaluation protocol, unfortunately, present subtle yet critical caveats, thereby resulting in a highly biased measure of robustness. To mitigate these issues, we construct a diagnostic dataset through counterfactual simulations and revisit a collection of recent multi-agent forecasting models. Interestingly, we find that most recent models are already partially resilient to perturbations of non-causal agents but still struggle to capture indirect causal effects that involve mediator agents.

To further enhance the causal robustness of the learned representations, we propose a regularization approach that seeks to preserve the causal effect of each individual neighbor in an embedding space. Specifically, we devise two variants that exploit annotations with different levels of granularity: (i) a contrastive-based regularizer using binary annotations of causal/non-causal agents; (ii) a ranking-based regularizer using continuous annotations of causal effects. Through controlled experiments, we show that both regularizers can enhance causal awareness to notable degrees. More crucially, we find that finer-grained annotations are particularly important for generalization out of the training distribution, such as higher agent density or unseen context arrangements.

Finally, we introduce a sim-to-real causal transfer framework, aiming at extending the strengths of causal regularization from simulation to real-world contexts. We achieve this through cross-domain multi-task learning, *i.e.*, jointly train the representation on the causal task in simulation and the forecasting task in the real world. Through experiments on the ETH-UCY dataset [34, 50] paired with an ORCA simulator [62], we find that the causal transfer framework enables stronger generalization in challenging settings such as low-data regimes, even in the absence of real-world causal annotations. As one of the first steps towards causal models in the multi-agent, we hope our work brings new light on the challenges and opportunities of learning causally-aware representations in practice.

2 Method

In this section, we will introduce a simple yet effective approach to promote causal awareness of the learned representations. We will first describe a general regularization method, which encapsulates two specific instances exploiting causal annotations of different granularity. We will then introduce a sim-to-real transfer framework that extends the regularization method to practical settings, even in the absence of real-world annotations. More detailed formalisms of causal relationships between interactive agents and pitfalls of existing benchmarks are referred to Appendix B.

2.1 Causally-Aware Regularization

Recall that the causal effect $\mathcal{E}_i$ of an agent i is tied to the *difference* of the potential outcomes between the factual scene and the counterfactual one where the agent i is removed. In this light, a causally-aware representation should also capture such relations, *i.e.*, feature vectors of paired scenes are separated by a certain distance d_i depending on $\mathcal{E}_i$. Motivated by this intuition, we measure the distance between a pair of counterfactual scenes in an embedding space through cosine distance,

$$d_i = 1 - \mathrm{sim}(\boldsymbol{p}_{\varnothing}, \boldsymbol{p}_i) = 1 - \boldsymbol{p}_{\varnothing}^{\top}\boldsymbol{p}_i / \|\boldsymbol{p}_{\varnothing}\|\|\boldsymbol{p}_i\|, \tag{1}$$

where $\boldsymbol{p}_i = h(\boldsymbol{z}_i) = h(f(\boldsymbol{x}_i))$ is a low-dimensional feature vector projected from the latent representation $\boldsymbol{z}_i$ through a non-linear head h. The desired representation is thus expected to preserve the following property,

$$\mathcal{E}_i < \mathcal{E}_j \implies d_i < d_j, \quad \forall i, j \in \mathcal{A}, \tag{2}$$

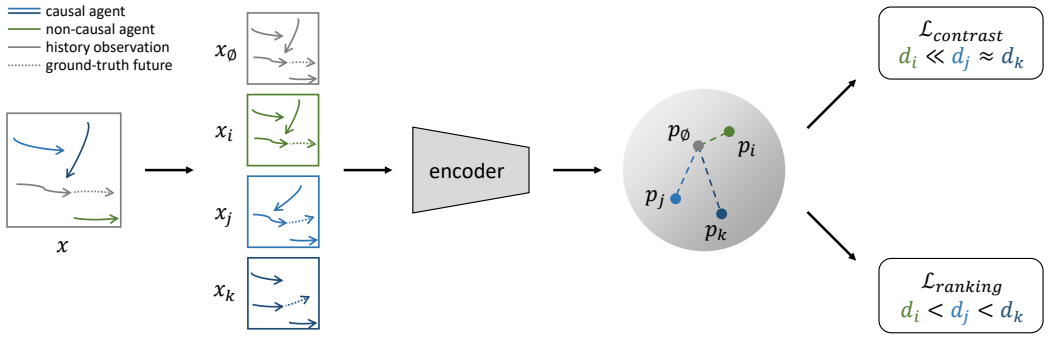


Figure 2: Overview of our method. We seek to build an encoder that captures the causal effect of each agent by regularizing the distance of paired embeddings between the factual and counterfactual scenes. We formulate this objective into a contrastive task given binary annotations or a ranking task given real-value annotations.

where i and j are the indices of two agents in the set of neighboring agents $\mathcal{A}$. We next describe two specific tasks that seek to enforce this property Eq. (2), while taking into account the concrete forms of causal annotations with different granularity, as illustrated in Fig. 2.

Causal Contrastive Learning. As discussed in Appendix B.3, one simple form of causal annotations is binary labels, indicating whether or not a neighboring agent causally influences the behavior of the ego agent. Intuitively, the representations of paired counterfactual scenes with respect to non-causal agents should be quite different from those with respect to causal ones: the former should stay close to the factual scenes in the embedding space, whereas the latter should be rather far away. We formulate this intuition into a causal contrastive learning objective,

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(d^+/\tau)}{\exp(d^+/\tau) + \sum_k \exp(d_k/\tau)\mathbb{1}_{\mathcal{E}_k > \eta}}, \tag{3}$$

where the positive (distant) example is sampled from counterfactual pairs with respect to causal agents, the negative (nearby) examples are sampled from counterfactual pairs with respect to non-causal agents, τ is a temperature hyperparameter controlling the difficulty of the contrastive task.

Causal Ranking Learning. One downside of causal contrastive learning described above is that it inherently ignores the detailed effect of causal agents. It tends to push the embeddings of counterfactuals equally far apart across all causal agents, regardless of the variation of causal effects, which violates the desired property stated in Eq. (2). To address this limitation, we further consider another variant of causal regularization using real-valued annotations to provide more dedicated supervision on the relative distance in the embedding space. Concretely, we first sort all agents in a scene based on their causal effect and then sample two agents with different causal effects for comparison. This allows us to formulate a ranking problem in a pairwise manner through a margin ranking loss,

$$\mathcal{L}_{\text{ranking}} = \max(0, d_i - d_j + m), \tag{4}$$

where d_i and d_j are the embedding distances with respect to two agents of different causal effects $\mathcal{E}_i < \mathcal{E}_j$, and m is a small margin hyperparameter controlling the difficulty of the ranking task.

2.2 Sim-to-real Causal Transfer

The causal regularization method described above relies upon the premise that annotations of causal effects are readily available. However, as elaborated in Appendix B.3, procuring such causal annotations in real-world scenarios can be highly difficult. To bridge this gap, we extend our causal regularization approach to a sim-to-real transfer learning framework. Our key idea is that, despite discrepancies between simulation environments and the real world, some underlying causal mechanisms like collision avoidance and group coordination are likely stable across domains. As such, in the absence of real-world causal annotations, we can instead leverage the causal annotations derived from the simulation counterparts to jointly train the model on the prediction task $\mathcal{L}_{task}^{real}$ on the real-world data and the causal distance task on the simulation data $\mathcal{L}_{causal}^{syn}$,

$$\mathcal{L} = \mathcal{L}_{\text{task}}^{\text{real}} + \alpha \mathcal{L}_{\text{causal}}^{\text{syn}}, \tag{5}$$

Table 1: Performance of modern representations on the created diagnostic dataset. The prediction errors and causal errors are generally substantial across all evaluated models. However, the errors associated with ACE-NC are rather marginal, suggesting that recent models are already partially robust to non-causal perturbations.

	ADE ↓	FDE ↓	ACE-NC ↓	ACE-DC ↓	ACE-IC ↓
D-LSTM [30]	0.329	0.677	0.027	0.532	0.614
S-LSTM [2]	0.314	0.627	0.031	0.463	0.523
Trajectron++ [58]	0.312	0.630	0.024	0.479	0.568
STGCNN [47]	0.307	0.564	0.049	0.330	0.354
AutoBots [20]	0.255	0.497	0.045	0.595	0.616

where α is a hyperparameter controlling the emphasis on the causal task. Despite its simplicity, we will show in §3.3 that the sim-to-real causal transfer framework effectively translates the causal knowledge absorbed from simulations to real-world benchmarks, *e.g.*, the ETH-UCY dataset [34, 50], even without any causal annotations from the latter.

3 Experiments

In this section, we will present a set of experiments to answer the following four questions:

1. How well do recent representations capture the causal relations between interactive agents?
2. Is our proposed method effective for addressing the limitations of recent representations?
3. Do finer-grained annotations provide any benefits for learning causally-aware representations?
4. Finally, does greater causal awareness offer any practical advantages in challenging scenarios?

Throughout our experiments, the multi-agent forecasting task is defined as predicting the future trajectory of the ego agent for 12 time steps, given the history observations of all agents in a scene in the past 8 time steps. More detailed experimental settings are summarized in D.

3.1 Robustness of Recent Representations

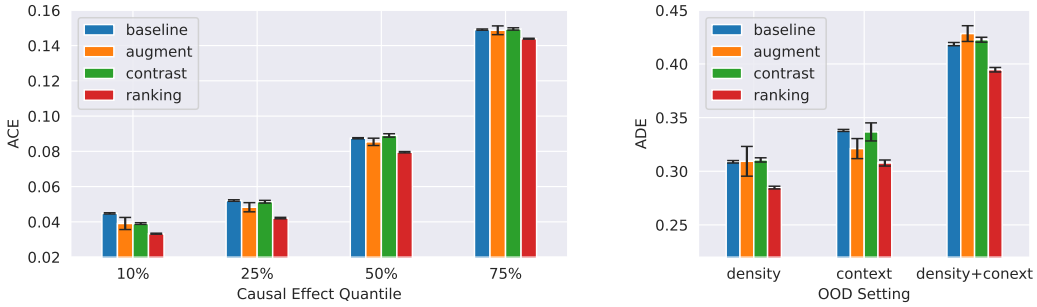
Setup. We start our experiments with the evaluation of a collection of recent models on the diagnostic dataset described in Appendix B, including S-LSTM [2], D-LSTM [30], Trajectron++ [58], STGCNN [47], and AutoBots [20]. Thanks to the fine-grained annotations from counterfactual simulations, we explicitly examine the errors of causal effect estimation for each agent category. We summarize the results in Tab. 1 and visualize the detailed estimation in Fig. 6.

Takeaway 1: recent representations are already partially robust to non-causal agent removal. As shown in Tab. 1, the values of ACE-NC are generally quite minimal compared to the results on other metrics. Across all evaluated models, the errors made upon non-causal agents are 10x smaller than that for causal agents (ACE-DC/IC). Corroborate with our analysis in Appendix B, this result provides a counterargument to the robustness issue benchmarked in Roelofs et al. [54], suggesting the importance of studying the robustness under causal perturbations as opposed to non-causal ones.

Takeaway 2: recent representations underestimate causal effects, particularly indirect ones. As shown in Fig. 6, the estimate of causal effect often deviates from the ground truth value. In particular, the learned representation tends to severely underestimate the influence of indirect causal agents. This result underscores the limitation of existing interaction representations in reasoning about a chain of causal relations transmitting across multiple agents.

3.2 Effectiveness of Causal Regularization

We next evaluate the efficacy of our causal regularization method on Autobots, the strongest baselines in Tab. 1. We take the data augmentation strategy proposed in [54] as a baseline, and compare different methods in two aspects: in-distribution causal awareness and out-of-distribution generalization.



(a) In-Distribution Evaluation

(b) Out-of-Distribution Evaluation

Figure 3: Quantitaive results of our causal regularization method in comparison to the Autobots baseline [20] and the non-causal data augmentation [54] Models trained by our ranking-based method yield enhanced accuracy in estimating causal effects and more reliable predictions on out-of-distribution test sets. Results are averaged over five random seeds.

3.2.1 In-distribution Causal Awareness

Setup. We first examine the efficacy of our method by measuring the ACE on the in-distribution test set. Since the error of the baseline model varies substantially across different ranges of causal effects (§3.1), we split the test set into four quantile segments. The result is summarized in Fig. 3.

Takeaway: our method boosts causal awareness thanks to fine-grained annotations. As shown in Fig. 3a, both the contrastive and ranking variants of our causal regularization approach result in lower causal errors than the vanilla baseline. In particular, the causal ranking method demonstrates substantial advantages over the other counterparts across all quantities of causal effects, confirming the promise of incorporating fine-grained annotations in learning causally-aware representations.

3.2.2 Out-of-Distribution Generalization

Setup. We further examine the efficacy of our method by measuring the prediction accuracy on out-of-distribution test sets. We consider three common types of distribution shifts in the multi-agent setting: higher density, unseen context, and both combined. We summarize the results of different methods on the OOD test sets in Fig. 3b and visualize the difference of prediction output in Fig. 8.

Takeaway: causally-aware representations lead to stronger out-of-distribution generalization. As shown in Fig. 3b, the prediction error from our causal ranking method is generally lower than that of the other methods on the OOD sets. In fact, the overall patterns between Fig. 3a and Fig. 3b are highly similar, indicating a substantial correlation between causal awareness and out-of-distribution robustness. This practical benefit is also visually evident in the qualitative visualization in Fig. 8.

3.3 Effectiveness of Causal Transfer

Setup. Finally, we evaluate the proposed causal transfer approach on the ETH-UCY dataset [34, 50] paired with our diagnostic dataset. To assess the potential of the transfer method in challenging settings such as low-data regimes, we train Autobots [20] on different fractions of the real-world data. The results are summarized in Fig. 7.

Takeaway: causal transfer improves performance in the real world despite domain gaps. As shown in Fig. 7, our causal transfer approach consistently improves the performance of the learned representation on the real-world benchmark. In particular, it enables the model to learn faster, *e.g.*, even with half data in the real-world, our ranking-based method can still outperform other counterparts. In contrast, the vanilla sim-to-real technique [36] fails to yield effective transfer, especially in the low-data regimes. We conjecture that this is due to the disparity between simulation and the real world in terms of non-causal interaction styles, such as speed and curvature.

Additional evaluations and discussions are summarized in Appendices C and E.

References

[1] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-Aware Large-Scale Crowd Forecasting. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2211–2218, June 2014.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, June 2016.

[3] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, Mar. 2020.

[4] S. S. G. Bagi, Z. Gharaee, O. Schulte, and M. Crowley. Generative Causal Representation Learning for Out-of-Distribution Motion Forecasting, Feb. 2023.

[5] Y. Cao, C. Xiao, A. Anandkumar, D. Xu, and M. Pavone. AdvDO: Realistic Adversarial Attacks for Trajectory Prediction. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 36–52, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20065-6.

[6] Y. Cao, D. Xu, X. Weng, Z. Mao, A. Anandkumar, C. Xiao, and M. Pavone. Robust Trajectory Prediction against Adversarial Attacks. In *Proceedings of The 6th Conference on Robot Learning*, pages 128–137. PMLR, Mar. 2023.

[7] L. Castri, S. Mghames, M. Hanheide, and N. Bellotto. Causal Discovery of Dynamic Models for Predicting Human Spatial Interactions. In F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, and S. S. Ge, editors, *Social Robotics*, Lecture Notes in Computer Science, pages 154–164, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-24667-8.

[8] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov. MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction. In *Conference on Robot Learning*, pages 86–99. PMLR, May 2020.

[9] C. Chen, Y. Liu, S. Kreiss, and A. Alahi. Crowd-Robot Interaction: Crowd-Aware Robot Navigation With Attention-Based Deep Reinforcement Learning. In *International Conference on Robotics and Automation (ICRA)*, pages 6015–6022, May 2019.

[10] G. Chen, J. Li, J. Lu, and J. Zhou. Human Trajectory Prediction via Counterfactual Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9824–9833, 2021.

[11] G. Chen, F. Liu, Z. Meng, and S. Liang. Revisiting Parameter-Efficient Tuning: Are We Really There Yet? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2612–2626, Dec. 2022.

[12] G. Chen, Z. Chen, S. Fan, and K. Zhang. Unsupervised Sampling Promoting for Stochastic Human Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17874–17884, 2023.

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016.

[14] C. Choi, S. Malla, A. Patil, and J. H. Choi. DROGON: A Trajectory Prediction Model based on Intention-Conditioned Behavior Reasoning. In *Proceedings of the 2020 Conference on Robot Learning*, pages 49–63. PMLR, Oct. 2021.

[15] N. Deo and M. M. Trivedi. Convolutional Social Pooling for Vehicle Trajectory Prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1549–15498, June 2018.

[16] A. Dittadi, F. Träuble, F. Locatello, M. Wuthrich, V. Agrawal, O. Winther, S. Bauer, and B. Schölkopf. On the Transfer of Disentangled Representations in Realistic Settings. In *International Conference on Learning Representations*, Sept. 2020.

[17] A. Dittadi, S. S. Papa, M. D. Vita, B. Schölkopf, O. Winther, and F. Locatello. Generalization and Robustness Implications in Object-Centric Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5221–5285. PMLR, June 2022.

[18] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[19] G. Ferrer, A. Garrell, and A. Sanfeliu. Robot companion: A social-force based approach with human awareness-navigation in crowded environments. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1688–1694, Nov. 2013.

[20] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D'Souza, S. E. Kahou, F. Heide, and C. Pal. Latent Variable Sequential Set Transformers for Joint Multi-Agent Motion Prediction. In *International Conference on Learning Representations*, Oct. 2021.

[21] J. Gu, Q. Sun, and H. Zhao. DenseTNT: Waymo Open Dataset Motion Prediction Challenge 1st Place Solution. *arXiv:2106.14160 [cs]*, June 2021.

[22] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, June 2018.

[23] D. Helbing and P. Molnar. Social Force Model for Pedestrian Dynamics. *Physics Review E*, May 1998.

[24] D. Hendrycks*, N. Mu*, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*, Mar. 2020.

[25] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

[26] Y. Hu, X. Jia, M. Tomizuka, and W. Zhan. Causal-based Time Series Domain Generalization for Vehicle Intention Prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7806–7813, May 2022.

[27] B. Huang, F. Feng, C. Lu, S. Magliacane, and K. Zhang. AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning. In *International Conference on Learning Representations*, Jan. 2022.

[28] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6272–6281, 2019.

[29] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks. *Advances in Neural Information Processing Systems*, 32:137–146, 2019.

[30] P. Kothari, S. Kreiss, and A. Alahi. Human Trajectory Forecasting in Crowds: A Deep Learning Perspective. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2021.

[31] P. Kothari, D. Li, Y. Liu, and A. Alahi. Motion Style Transfer: Modular Low-Rank Adaptation for Deep Motion Forecasting. In *Conference on Robot Learning (CoRL)*, Nov. 2022.

[32] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826. PMLR, July 2021.

[33] Y. Lee, A. S. Chen, F. Tajwar, A. Kumar, H. Yao, P. Liang, and C. Finn. Surgical Fine-Tuning Improves Adaptation to Distribution Shifts. In *International Conference on Learning Representations*, 2023.

[34] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by Example. *Computer Graphics Forum*, 26:655–664, 2007.

[35] J. Li, F. Yang, M. Tomizuka, and C. Choi. EvolveGraph: Multi-Agent Trajectory Prediction with Dynamic Relational Reasoning. In *Advances in Neural Information Processing Systems*, volume 33, pages 19783–19794. Curran Associates, Inc., 2020.

[36] J. Liang, L. Jiang, and A. Hauptmann. SimAug: Learning Robust Representations from Simulation for Trajectory Prediction. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 275–292, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58601-0.

[37] Y. Liu, Q. Yan, and A. Alahi. Social NCE: Contrastive Learning of Socially-Aware Motion Representations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15118–15129, 2021.

[38] Y. Liu, R. Cadei, J. Schweizer, S. Bahmani, and A. Alahi. Towards Robust and Adaptive Motion Forecasting: A Causal Representation Perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17081–17092, 2022.

[39] Y. Liu, A. Alahi, C. Russell, M. Horn, D. Zietlow, B. Schölkopf, and F. Locatello. Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning. In *Conference on Causal Learning and Reasoning (CLeaR)*, Apr. 2023.

[40] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, May 2019.

[41] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. Object-Centric Learning with Slot Attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc., 2020.

[42] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People tracking with human motion predictions from social forces. In *2010 IEEE International Conference on Robotics and Automation*, pages 464–469, May 2010.

[43] O. Makansi, J. V. Kügelgen, F. Locatello, P. V. Gehler, D. Janzing, T. Brox, and B. Schölkopf. You Mostly Walk Alone: Analyzing Feature Attribution in Trajectory Prediction. In *International Conference on Learning Representations*, Sept. 2021.

[44] K. Mangalam, Y. An, H. Girase, and J. Malik. From Goals, Waypoints & Paths to Long Term Human Trajectory Forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15233–15242, 2021.

[45] D. McDuff, Y. Song, J. Lee, V. Vineet, S. Vemprala, N. A. Gyde, H. Salman, S. Ma, K. Sohn, and A. Kapoor. CausalCity: Complex Simulations with Agency for Causal Discovery and Reasoning. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 559–575. PMLR, June 2022.

[46] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, June 2009.

[47] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14412–14420, June 2020.

[48] M. L. Montero, C. J. Ludwig, R. P. Costa, G. Malhotra, and J. Bowers. The role of Disentanglement in Generalisation. In *International Conference on Learning Representations*, Feb. 2022.

[49] H.-S. Moon and J. Seo. Fast User Adaptation for Human Motion Prediction in Physical Human–Robot Interaction. *IEEE Robotics and Automation Letters*, 7:120–127, Jan. 2022.

[50] S. Pellegrini, A. Ess, and L. Van Gool. Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, pages 452–465, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-15549-9.

[51] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, USA, Nov. 2017. ISBN 978-0-262-03731-0.

[52] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400. PMLR, May 2019.

[53] N. Rhinehart, R. Mcallister, K. Kitani, and S. Levine. PRECOG: PREdiction Conditioned on Goals in Visual Multi-Agent Settings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2821–2830, Oct. 2019.

[54] R. Roelofs, L. Sun, B. Caine, K. S. Refaat, B. Sapp, S. Ettinger, and W. Chai. CausalAgents: A Robustness Benchmark for Motion Forecasting using Causal Relationships, Oct. 2022.

[55] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39: 895–935, July 2020.

[56] S. Saadatnejad, M. Bahari, P. Khorsandi, M. Saneian, S.-M. Moosavi-Dezfooli, and A. Alahi. Are socially-aware trajectory prediction models really socially-aware? *Transportation Research Part C: Emerging Technologies*, 141:103705, Aug. 2022.

[57] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[58] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 683–700, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58523-5.

[59] B. Schölkopf. Causality for Machine Learning. *arXiv:1911.10500 [cs, stat]*, Dec. 2019.

[60] B. Schölkopf and J. von Kügelgen. From Statistical to Causal Learning. *arXiv:2204.00607 [cs, stat]*, Apr. 2022.

[61] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109:612–634, May 2021.

[62] J. van den Berg, M. Lin, and D. Manocha. Reciprocal Velocity Obstacles for real-time multi-agent navigation. In *2008 IEEE International Conference on Robotics and Automation*, pages 1928–1935, May 2008.

[63] J. van den Berg, S. J. Guy, M. Lin, and D. Manocha. Reciprocal n-body collision avoidance. In C. Pradalier, R. Siegwart, and G. Hirzinger, editors, *Robotics Research*, pages 3–19, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19457-3.

[64] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are Disentangled Representations Helpful for Abstract Visual Reasoning? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[65] A. Vemula, K. Muelling, and J. Oh. Social Attention: Modeling Attention in Human Crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7, Brisbane, Australia, May 2018. IEEE Press.

[66] C. Xu, T. Li, C. Tang, L. Sun, K. Keutzer, M. Tomizuka, A. Fathi, and W. Zhan. PreTraM: Self-supervised Pre-training via Connecting Trajectory and Map. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 34–50, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19842-7.

[67] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang. EqMotion: Equivariant Multi-Agent Motion Prediction With Invariant Interaction Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023.

[68] F. Zanlungo, T. Ikeda, and T. Kanda. Social force model with explicit collision prediction. *EPL (Europhysics Letters)*, 93:68005, Mar. 2011.

[69] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, and M. Tomizuka. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps, Sept. 2019.

[70] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto, C. Finn, and A. Gupta. Train Offline, Test Online: A Real Robot Learning Benchmark, June 2023.

A Related Work

The social causality studied in this work lies at the intersection of three areas: multi-agent interactions, robust representations, and causal learning. In this section, we provide a brief overview of the existing literature in each area and then discuss their relevance to our work.

Multi-Agent Interactions. The study of multi-agent interactions has a long history. Early efforts were focused on hand-crafted rules, such as social forces [23, 46] and reciprocal collision avoidance [1, 62]. Despite remarkable results in sparse scenarios [19, 42, 68], these models often lack social awareness in more densely populated and complex environments [55]. As an alternative, recent years have witnessed a paradigm shift toward learning-based approaches, particularly the use of carefully designed neural networks to learn representations of multi-agent interactions [30]. Examples include pooling operators [2, 15, 22], attention mechanisms [28, 57, 65], spatio-temporal graphs [29, 35, 58], among others [8, 14, 53]. Nevertheless, the robustness of these models remains a grand challenge [6, 54, 56]. Our work presents a solution to enhance robustness by effectively exploiting causal annotations of varying granularity.

Robust Representations. The robustness of machine learning models, especially under distribution shifts, has been a long-standing concern for safety-critical applications [52]. Existing efforts have explored two main avenues to address this challenge. One line of work seeks to identify features that are invariant across distributions. Unfortunately, this approach often relies on strong assumptions about the underlying shifts [24, 37] or on access to multiple training domains [3, 32], which may not be practical in real-world settings. Another approach aims to develop models that can efficiently adapt to new distributions by updating only a small number of weight parameters, such as sub-modules [31], certain layers [33], or a small subset of neurons [11]. More recently, there has been a growing interest in exploring the potential of causal learning to address the robustness challenges [16, 17, 48, 64]. To the best of our knowledge, our work makes the first attempt in the multi-agent context, showcasing the benefits of incorporating causal relationships for stronger out-of-distribution generalization in more crowded spaces.

Causal Learning. Empowering machine learning with causal reasoning has gained a growing interest in recent years [59, 60]. One line of work seeks to discover high-level causal variables from low-level observations, *e.g.*, disentangled [13, 25, 40] or structured latent representations [41, 61]. Unfortunately, existing methods remain largely limited to simple and static datasets [39]. Another thread of work attempts to draw causal insights into dynamic decision-making [27]. In particular, recent works have proposed a couple of methods to incorporate causal invariance and structure into the design and training of forecasting models in the multi-agent context [4, 7, 10, 26, 38, 43]. However, these efforts have mainly focused on using causal implications to enhance robustness rather than explicitly examining causal relations between interactive agents. Closely related to ours, another recent study introduces a motion forecasting benchmark with annotations of causal relationships [54]. Our work takes a critical look at its labeling as well as evaluation protocols and designs practical methods to enhance causal awareness.

B Formalism

In this section, we seek to formalize the robustness challenge in multi-agent representation learning through the lens of social causality. We will first revisit the design of the CausalAgents [54] benchmark, drawing attention to its caveats in labeling and evaluation. Subsequently, we will introduce a diagnostic dataset, aiming to facilitate more rigorous development and evaluation of causally-aware representations through counterfactual simulations.

B.1 Social Interaction and Social Causality

Social Interaction. Consider a motion forecasting problem where an ego agent is surrounded by a set of neighboring agents in a scene. Let $s_t^i = (x_t^i, y_t^i)$ denote the state of agent i at time t and $s_t = \{s_t^0, s_t^1, \cdots, s_t^K\}$ denote the joint state of all the agents in the scene. Without loss of generality, we index the ego agent as 0 and the rest of neighboring agents as $\mathcal{A} = \{1, 2, \cdots, K\}$. Given a sequence of history observations $\boldsymbol{x} = (s_1, \cdots, s_t)$, the task is to predict the future trajectory of the

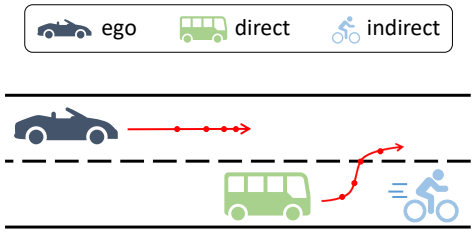


Figure 4: Illustration of indirect causal effects. The cyclist *indirectly* influences the decision of the ego agent due to the presence of the bus.

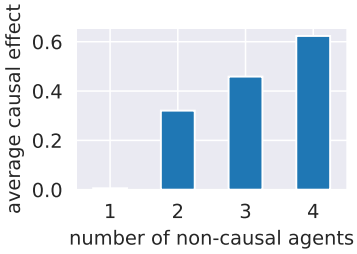


Figure 5: Joint effect of all non-causal agents in the scene, simulated in ORCA. The ego agent often changes its behavior when 2+ non-causal neighbors are removed.

ego agent $\boldsymbol{y} = (s^0_{t+1}, \cdots, s^0_T)$ until time T. Modern forecasting models are largely composed of encoder-decoder neural networks, where the encoder $f(\cdot)$ first extracts a compact representation $\boldsymbol{z}$ of the input with respect to the ego agent and the decoder $g(\cdot)$ subsequently rolls out a sequence of its future trajectory $\hat{\boldsymbol{y}}$:

$$\begin{aligned} \boldsymbol{z} &= f(\boldsymbol{x}) = f(s_{1:t}), \\ \hat{\boldsymbol{y}} &= \hat{s}^0_{t+1:T} = g(\boldsymbol{z}). \end{aligned} \tag{6}$$

Social Causality. Despite remarkable progress on accuracy measures [2, 58], recent neural representations of social interactions still suffer from a significant robustness concern. For instance, recent works have shown that trajectories predicted by existing models often output colliding trajectories [37], vulnerable to adversarial perturbations [5, 56] and deteriorate under distribution shifts of spurious feature such as agent density [38].

One particular notion of robustness that has recently gained much attention is related to the causal relationships between interactive agents [54]. Ideally, neural representations of social interactions should capture the influence of neighboring agents, namely how the future trajectory of the ego agent will vary between a pair of scenes: an original scene $\boldsymbol{x}_\varnothing$ and a perturbed scene $\boldsymbol{x}_R$ where some neighboring agents $\mathcal{R}$ are removed, *i.e.*, only the ego agent and $\mathcal{A} \setminus \mathcal{R}$ remain. We define the causal effect of the removed agents $\mathcal{R}$ as,

$$\mathcal{E}_\mathcal{R} = \|\boldsymbol{y}_\varnothing - \boldsymbol{y}_\mathcal{R}\|_2, \tag{7}$$

where $\boldsymbol{y}_\varnothing \equiv \boldsymbol{y}$ is the future trajectory in the original scene, $\boldsymbol{y}_\mathcal{R}$ is the future trajectory in the perturbed scene, and $\|\cdot\|_2$ is the average point-wise Euclidean distance between two trajectories.

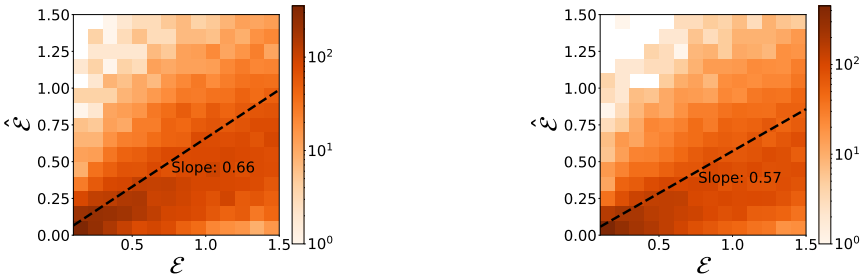
B.2 Caveats of Social Causality Benchmark

While the mathematical definition of the causal effect is straightforward in the multi-agent context, measuring Eq. (7) in practice can be highly difficult. In general, it is impossible to observe a subset of agents replaying their behaviors in the same environment twice in the real world – an issue known as the impossibility of counterfactuals [51]. To mitigate the data collection difficulty, a recent benchmark CausalAgents [54] proposes another simplified labeling strategy: instead of collecting paired counterfactual scenes, it queries human labelers to divide neighboring agents into two categories: causal agents that directly influence the driving behavior of the ego agent from camera viewpoints; and non-causal agents that do not. This labeling approach is accompanied by an evaluation protocol through agent removal, assuming that robust forecasting models should be insensitive to scene-level perturbations that remove non-causal agents. More formally, the CausalAgents benchmark evaluates the robustness of a learned representation through the following measure,

$$\Delta = \|\hat{\boldsymbol{y}}_\mathcal{R} - \boldsymbol{y}_\mathcal{R}\|_2 - \|\hat{\boldsymbol{y}}_\varnothing - \boldsymbol{y}_\varnothing\|_2, \tag{8}$$

where $\mathcal{R}$ is the set of all (or some) non-causal agents in the scene.

Caveats. In spite of the precious efforts on human annotations, the CausalAgents benchmark [54] is unfortunately plagued by two subtle but fundamental flaws.



(a) Direct Causal Agents **(b)** Indirect Causal Agents

Figure 6: Comparison between estimated causal effects $\hat{\mathcal{E}}$ and the corresponding ground truth $\mathcal{E}$. We uniformly sample the ground-truth causal effect, collect the estimate from AutoBots [20], and linearly regress a slope between the estimated value and the ground truth. The regression slope for a perfect model is 1.

1. *Annotation*: The labeling rule completely overlooks indirect causal relationships, *i.e.*, a neighbor does not directly influence the behavior of the ego agent but does so indirectly by influencing one or a few other neighbors that pass the influence to the ego agent, as illustrated in Fig. 4. On the one hand, indirect causal effects are practically non-trivial for human labelers to identify due to complex relation chains. On the other hand, they are prevalent in densely populated scenes, not only posing a significant modeling challenge but also playing a crucial role in causal learning, which we will demonstrate in §3.
2. *Evaluation*: The evaluation protocol tends to dramatically overestimate the robustness issue. As per Eq. (8), non-causal agents are delineated at the individual level, but robustness assessment occurs at the category level, *e.g.*, removing all non-causal agents rather than a single one. In fact, the joint effect of non-causal agents can escalate quickly with an increasing number of agents, as demonstrated in Fig. 5. Such a discrepancy between annotation and evaluation with respect to non-causal agents can lead to an inflated perception of the robustness issue.

B.3 Diagnostic Dataset through Counterfactuals

Counterfactual Pairs. To address the above caveats, we create a new diagnostic dataset through counterfactual simulations with ORCA [62]. Specifically, we annotate the ground-truth causal effect by explicitly comparing the trajectories in paired scenes before and after agent removals, as defined in Eq. (7). Details of the collected data are summarized in Appendix D.

Fine-grained Category. In addition to real-valued causal effects, we also seek to annotate the category of each agent. Specifically, from simulations, we extract the per-step causal relation between a neighboring agent i and the ego agent. If agent i is visible to the ego at time step t, it directly influences the ego, indicated by $\mathbb{1}_t^i = 1$; otherwise $\mathbb{1}_t^i = 0$. We then convert the causal effect over the whole sequence $\mathcal{E}_i$ into three agent categories:

- *Non-causal agent*: little influence on the behavior of the ego agent, *i.e.*, $\mathcal{E} < \epsilon \approx 0$.
- *Direct causal agent*: significant influence on the behavior of the ego agent $\mathcal{E} > \eta \gg 0$; moreover, the influence is direct for at least one time step $\prod_{\tau=1:T}(1 - \mathbb{1}_\tau^i) = 0$.
- *Indirect causal agent*: significant influence on the behavior of the ego agent $\mathcal{E} > \eta \gg 0$; however, the influence is never direct over the entire sequence $\prod_{\tau=1:T}(1 - \mathbb{1}_\tau^i) = 1$.

The collected diagnostic dataset with different levels of causal annotations allows us to rigorously probe the causal robustness of existing representations and to develop more causally-aware representations, which we will describe in the next sections.

C Additional Results

In addition to the aggregated results presented in Fig. 7, we summarize an in-depth breakdown of the sim-to-real transfer results in Tabs. 2 to 4. Across all evaluated settings, our ranking-based causal transfer consistently achieves superior prediction accuracy compared to the AutoBots baseline. Notably, it outpaces the standard sim-to-real method [36] in four out of the five subsets, with the

Table 2: Quantitative results of sim-to-real transfer from ORCA simulations to 25% of the ETH-UCY dataset.

25%	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
AutoBots [20]	0.942/1.886	0.333/0.662	0.563/1.156	0.446/0.958	0.409/0.904	0.539/1.113
+ Vanilla [36]	0.956/1.893	0.352/0.698	**0.537/1.143**	0.439/0.969	0.426/0.935	0.542/1.128
+ Augment [54]	0.942/1.867	0.359/0.722	0.544/1.152	0.441/0.971	0.409/0.922	0.539/1.127
+ Contrast (ours)	0.938/1.885	**0.320/0.616**	0.565/1.187	0.450/0.971	0.386/0.821	0.532/1.096
+ Ranking (ours)	**0.920/1.836**	0.326/0.654	0.555/1.144	**0.438/0.945**	**0.376/0.824**	**0.523/1.081**

Table 3: Quantitative results of sim-to-real transfer from ORCA simulations to 50% of the ETH-UCY dataset.

50%	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
AutoBots [20]	0.940/1.883	0.326/0.612	0.566/1.197	0.434/0.945	0.358/0.787	0.525/1.085
+ Vanilla [36]	0.923/1.913	0.342/0.661	**0.535/1.141**	0.430/0.954	0.383/0.886	0.523/1.111
+ Augment [54]	0.937/1.885	0.340/0.660	0.535/1.146	0.424/0.938	0.377/0.878	0.523/1.101
+ Contrast (ours)	0.935/1.870	0.344/0.667	0.554/1.148	**0.422/0.913**	0.346/0.772	0.520/1.074
+ Ranking (ours)	**0.915/1.826**	**0.303/0.567**	0.549/1.159	0.427/0.919	**0.340/0.748**	**0.507/1.044**

Table 4: Quantitative results of sim-to-real transfer from ORCA simulations to 100% of the ETH-UCY dataset.

100%	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
AutoBots [20]	0.938/1.916	0.334/0.678	0.550/1.152	0.420/0.916	0.343/0.787	0.517/1.090
+ Vanilla [36]	0.923/1.913	0.331/0.638	0.527/1.133	0.410/0.907	0.346/0.783	0.507/1.075
+ Augment [54]	0.938/1.878	0.332/0.635	**0.518/1.141**	**0.403/0.890**	0.348/0.796	0.508/1.068
+ Contrast (ours)	0.930/1.896	0.321/0.605	0.538/1.154	0.412/0.899	0.345/0.791	0.509/1.069
+ Ranking (ours)	**0.916/1.846**	**0.312/0.582**	0.539/1.145	0.408/0.891	**0.326/0.714**	**0.500/1.036**

sole exception in the UNIV subset. We conjecture that this exception might be attributed to the high similarity between the ORCA simulation and UNIV dataset.

Furthermore, as summarized in Tab. 5, AutoBots stands as one of the state-of-the-art models for multi-agent trajectory forecasting, leaving only marginal room for improvement on the ETH-UCY dataset. In spite of this, our proposed causal transfer method still offers notable improvements, resulting in comparable accuracies to EqMotion [67], the most recent model that leverages domain-specific knowledge for predictive tasks.

D Implementation Details

Experiment details. Our experiments are largely built upon the public code of prior work, with as few modifications as possible made for the implementations of our proposed regularizers. Concretely, in the robustness analysis reported in Tab. 1, we train each model on our constructed dataset using the default hyperparameters of the corresponding baseline. To understand the performance of our proposed methods in §3.2, we fine-tune the pre-trained checkpoint for 10 epochs, and evaluate the obtained model on the hold-out test set. The main hyperparameters used for training our baseline model AutoBots [20] and the causal regularizers are listed in Tab. 6.

Dataset details. Our diagnostic dataset is generated using a customized version of the Reciprocal Velocity Obstacle simulator that employs Optimal Reciprocal Collision Avoidance (ORCA) [63]. To simulate realistic causal relationships between agents, we imposed a visibility constraint where an agent only observes other neighbors within their proximity and its 210° field of view. This visibility plays a significant role in determining the influence of one agent on another. Specifically, we define a neighbor i as having a *direct influence* on the ego agent at time step t if it is visible to the ego in that time step, *i.e.*, $\mathbb{1}_t^i = 1$. Additionally, we introduce a visibility window that records agents that were previously visible, facilitating the modeling of a richer spectrum of direct and indirect inter-agent influences. To encourage the presence of non-causal agents in dense spaces, we explicitly directed specific agents to follow others, thereby making them non-causal or indirect causal. Tab. 7 summarize the key statistics of our diagnostic datasets, for the original in-distribution dataset, as well as the out-of-distribution counterparts.

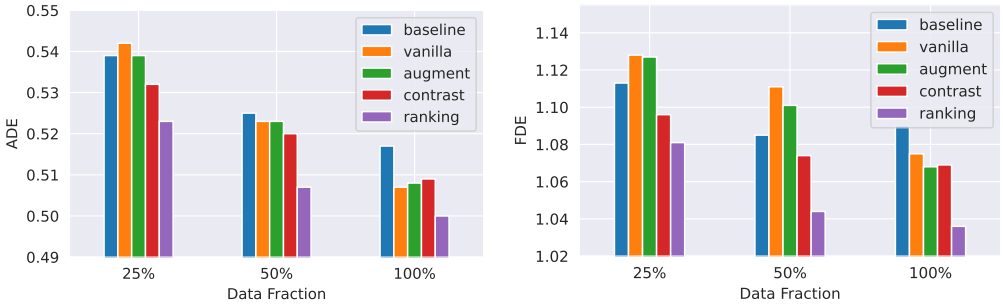


Figure 7: Results of causal transfer from ORCA simulations to the ETH-UCY dataset. We consider three settings where the model has access to the same simulation data but varying amounts of real-world data. Our ranking-based causal transfer results in lower prediction errors and higher learning efficiency than the other counterparts, *i.e.*, AutoBots baseline [20], the vanilla sim-to-real transfer [36] and the non-causal data augmentation [54]. Details are summarized in Appendices C and D

.

Table 5: Comparison between different multi-agent forecasting models on the ETH-UCY dataset. Boosted by our proposed ranking-based causal transfer, the best result of AutoBots across three random seeds reaches comparable performance to the current state-of-the-art.

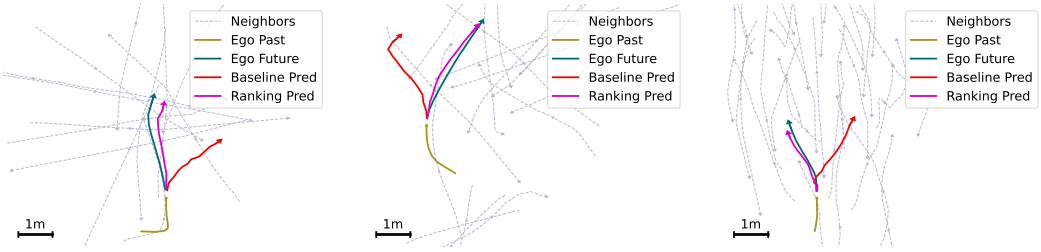
Deterministic	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
S-LSTM [2]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
D-LSTM [30]	1.05/2.10	0.46/0.93	0.57/1.25	0.40/0.90	0.37/0.89	0.57/1.21
Trajectron++ [58]	1.02/2.00	0.33/0.62	0.53/1.19	0.44/0.99	0.32/0.73	0.53/1.11
EqMotion[67]	0.96/1.92	**0.30**/0.58	**0.50/1.10**	**0.39/0.86**	**0.30/0.68**	**0.49**/1.03
AutoBots [20]	0.93/1.87	0.32/0.65	0.54/1.15	0.42/0.91	0.34/0.77	0.51/1.07
AutoBots + Ranking (ours)	**0.90/1.81**	**0.30/0.56**	0.53/1.12	0.41/0.89	0.32/0.71	**0.49/1.02**

Baseline details. To the best of our knowledge, there are few prior work that studies causal representation learning in the multi-agent context. The only one that explicitly leverages causal annotations is the data augmentation strategy proposed in the CausalAgents benchmark. We adopt their implementation practice in our experiment in §3.2, training the model from scratch on data that contains both factual and counterfactual scenes with respect to non-causal agents. To understand the efficacy of the proposed sim-to-real causal transfer, we implement the core element of the vanilla sim-to-real transfer [36] as the baseline. Specifically, we evenly blend the simulation and real-world data in each batch, training the model on the prediction task that spans both domains simultaneously.

E Additional Discussions

Summary. In this paper, we presented a thorough analysis and an effective approach for learning causally-aware representations of multi-agent interactions. We cast doubt on the notion of non-causal robustness in a recent benchmark [54] and showed that the main weaknesses of recent representations are not overestimations of the effect of non-causal agents but rather underestimation of the effects of indirect causal agents. To boost causal awareness of the learned representations, we introduced a regularization approach that encapsulates a contrastive and a ranking variant leveraging annotations of different granularity. We showed that our approach enables recent models to learn faster, generalize better, as well as transfer stronger to practical problems, even without real-world annotations.

Limitations. As one of the first steps towards causal representations in the multi-agent context, our work is subject to two major limitations. On the technical front, while our proposed regularization method consistently boosts causal awareness in various settings, supervision alone is likely insufficient to fully solve causal reasoning in complex scenes, as evidenced by the causal errors in Fig. 3a. Incorporating structural inductive biases might be a promising direction to address high-order reasoning challenges in the presence of indirect causal effects. On the empirical side, while we have demonstrated the potential of our approach in real-world settings in §3.3, most of our other experiments were conducted in controlled simulations for proof-of-concept. Scaling our findings



(a) ID **(b)** OOD-Density **(c)** OOD-Context

Figure 8: Qualitative results of our method on in-distribution (ID) and out-of-distribution (OOD) test sets. Models regularized by our ranking-based method demonstrate more robust understanding of agent interactions in scenarios with higher agent density and unseen scene context.

to other practical benchmarks (*e.g.*, WOMD [18] and INTERACTION [69]) and contexts (*e.g.*, navigation [9] and manipulation [70]) can be another fruitful avenue for future research.

Counterfactual simulation. Our diagnostic dataset, enabled by counterfactual simulations, offers clean annotations of causal relationships, serving as a crucial step in understanding causally-aware representation of multi-agent interactions. However, the realism of these simulated causal effects is still subject to some inherent limitations. For example, we have enforced a stringent constraint on the field of view for each agent, considering that the ego agent is usually unaffected by trailing neighbors. Such constraints could compromise the optimality of the ORCA algorithm, potentially resulting in unnatural trajectories. We believe that integrating more advanced simulators, *e.g.*, CausalCity [45], can address these challenges and we anticipate promising outcomes along this line for future research.

Multi-agent causal effects. Our annotation and evaluation have been focused on the causal effect at an individual agent level, namely we remove only one agent at a time. Through this lens, we observe that while recent representations are already partially robust to non-causal agent removal, they tend to underestimate the effects of causal agents. However, it is worth noting that this is still a rather simplified and restricted setting compared to the group-level causal effects, where the collective behavior of multiple agents may have a more complex influence on the ego agent. Understanding and addressing this challenge can be another exciting avenue for future research.

Table 6: Key hyper-parameters in our experiments.

name	value
batch size	16
pre-training learning rate	7.5×10^{-4}
fine-tuning learning rate	2.34375×10^{-5}
contrastive weight α	1000
ranking weight α	1000
ranking margin m	0.001
non-causal threshold ϵ	0.02
causal threshold η	0.1

Table 7: Key statistics of our diagnostic datasets.

Dataset	Context	Number of scenes		Number of agents per scene			
		train	test	non-causal	direct causal	indirect causal	total
ID	Open area	20k	2k	1.31	8.35	0.48	12.03
OOD-Density	Open area	-	2k	9.47	12.27	2.55	28.98
OOD-Context	Street	-	2k	8.23	12.22	3.09	29.01
OOD-Density+Context	Plaza	-	2k	7.49	14.64	2.78	28.93