Towards Robust and Adaptable Motion Forecasting: A Causal Representation Perspective

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning behavioral patterns from observational data has been a *de-facto* approach to motion forecasting. Yet, the current paradigm suffers from two fundamental shortcomings: brittle under covariate shift and inefficient for knowledge transfer. In this work, we propose to address these challenges from a causal representation perspective. We first introduce a causal formalism of motion forecasting, which casts the problem as a dynamic process with physical mechanisms, style confounders, and spurious correlations. We then propose two components that explicitly promote the robustness and reusability of the learned motion presentations: (i) unlike the common practice of merging datasets collected from different locations, we exploit their subtle distinctions by means of an invariance loss function, which encourages the model to suppress spurious correlations and capture physical mechanisms; (ii) we devise a modular architecture that factorizes the representations of physical laws and motion styles in a structured way, and progressively prune their dense connections during training to approximate a sparse causal graph. We empirically validate the strength of the proposed method for robust generalization in controlled real-world experiments. We finally discuss the challenges and opportunities in the presence of style shifts through synthetic simulations.

1 Introduction

Motion forecasting is essential for autonomous systems running in dynamic environments. Yet, it is a challenging task due to strong spatial-temporal interactions. These interactions typically arise from two sources: physical laws (*e.g.*, collision avoidance) that govern the general dynamics; social norms (*e.g.*, social distance, left or right-hand traffic) that influence the motion style. Classical models attempt to describe these interactions based on the knowledge of physics, but often fall short of social awareness in complex scenes [1, 2]. As an alternative, learning motion representations from observational data has become a *de-facto* approach [3, 4]. In light of the rapid progress, solving motion forecasting is seemingly just around the corner by pursuing this fashion at larger scales.

However, the promise of the current paradigm is shadowed by two dark clouds:

- Despite the large modeling capacity, existing neural motion models struggle to discover the physical laws from data and often produce inadmissible outputs (*e.g.*, colliding trajectories) under covariate shifts [5, 6].
- Compared with human cognition, existing neural motion models are exceedingly inefficient for knowledge transfer and often require a large number of observations to adapt from one environment to another, even if the underlying change is local and mild [7].

These issues do not become any less severe with larger models [8]. Instead, they are profoundly rooted in the principle of statistical learning that only seeks correlations for the prediction task at hand, regardless of their robustness and reusability across domains.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In this work, we aim to tackle these challenges from a causal perspective. Incorporating causal relations into statistical modeling has recently garnered growing interest, as it not only offers a mathematical language to articulate distribution changes [9, 10] but also brings critical insights to representation learning [11–13]. There are also evidences from cognitive science that the notion of causality is prevalent in the motion context: few-month-old infants are already able to reason sensibly about the physical and social causality [14]; they can even learn causal relations simply from observations of interactions between adults, without any hands-on experience of their own [15]. How can we build forecasting models capable of capturing the *causal mechanisms* behind human motions from observational data?

To this end, we first introduce a formalism of motion forecasting through the lens of causality. As shown in Figure 2, motion forecasting can be cast as a dynamical process with three groups of latent variables: (i) domain-invariant causal variables that account for the physical laws universal to everyone at any place, (ii) domain-specific confounders that are associated with the motion style, which may vary from site to site, (iii) non-causal spurious features, whose correlations with the future motion may change drastically under different conditions. For instance, when a pedestrian walks into a crowded spot, two phenomena may emerge concurrently: the observed trajectory becomes noisy due to the perception (*e.g.*, tracking) errors under occlusions; the future trajectory turns non-linear because of strong interactions with other neighbors. Given this coincidence, the level of noise can bring rich information to the motion forecasting task. However, models built upon the strong but non-causal cue can be highly brittle. Any software or hardware updates related to the perception algorithms may induce changes to the noise pattern and thereby hurt prediction accuracy.

To enhance the robustness of forecasting models, we propose to promote causal invariance of the learned motion representations by exploiting the commonalities across multiple domains. Oftentimes, the training dataset for motion forecasting is not from a single place but comprises multiple subsets collected from different locations. Previous work typically merges them into a larger one, *e.g.*, the notable leave-one-out strategy for the ETH-UCY datasets [16, 17]. While this convention enriches the diversity of training data, each subset is often inherently distinctive [18]. Directly combining them not only entails a risk of inconsistent motion styles in the constructed learning problem but also destroys the critical information that reveals the stability of correlations across environments. To mitigate this issue, we draw inspiration from the principle of invariant risk minimization [19–21] and train the forecasting model with a penalty on the gradient norm. This regularizer encourages the model to suppress the spurious features and extract causally invariant features only. As a consequence, the resulting model is expected to generalize well to novel scenarios.

One downside of invariance regularization is that it enforces the model to ignore the style confounders that differ across environments. Intuitively, neither learning an ‘average‘ style using the conventional approach nor completely ignoring it by means of invariant risk is desirable for motion forecasting. In fact, in the presence of domain-specific confounders, it is not even possible for a model to directly generalize out-of-distribution, yet it remains crucial for the model to quickly adapt from one environment to another given limited samples. To facilitate knowledge transfer, we introduce a modular architecture that factorizes the representations of physical laws and social norms in a structural way. Our core assumption is that motion behaviors do not completely differ between environments; instead, style changes may stem from only some factors at a time, *e.g.*, social distance is smaller in train stations than in national parks. Moreover, the motion style may not have the same level of effects as physical laws. Based on these assumptions, we devise a structured architecture that contains two encoders to account for the physical laws and social norms separately. We initialize connections between these two groups of representations through normalization layers of the decoder and progressively prune the inter-module edges during training to approximate a sparse causal graph [22]. This design greatly reduces the redundant interplay between independent mechanisms [23–26], and hence allows the model to precisely localize and adapt a small subset of parameters responsible for the underlying style shift.

We examine the proposed method in two settings: synthetic simulation datasets and controlled real-world experiments. Our preliminary results demonstrate that the motion forecasting model trained by our method can generalize better than other previous methods in the presence of spurious correlations. Our results also shed light on the risk of the invariant principle under style shifts, a phenomenon unique to motion problems and revealing the vital importance of adaptable forecasting models in the few-shot setting. We hope our findings will pave the way for causal motion forecasting, an under-explored yet promising paradigm towards reliable and adaptable autonomy.

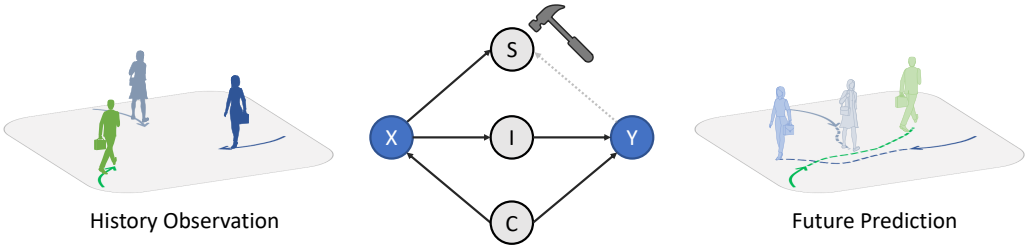

Figure 1: Our causal formalism of motion forecasting. We cast the problem as a dynamic process with three groups of latent variables: domain-invariant physical laws (I), domain-specific style confounders (C), and non-causal spurious features (S). The spurious features are not parents of future movements (Y) in the causal factorization, *e.g.*, no edges or anti-causal (dotted line), and their statistical correlations may vary drastically under different conditions. This formalism motivates our design and learning method to promote the robustness and reusability of the learned motion representation.

2 Method

Preliminary. Consider a motion forecasting problem in the multi-agent environments. For a scene of M agents, let $s_t = \{s_t^1, \cdots, s_t^M\}$ denote their joint state and $s_t^i = (x_t^i, y_t^i)$ denote the state of an individual agent i at time t. The task for the model is to take as input a sequence of past observations $x = (s_1, \cdots, s_{\bar{t}})$ and predict their states in the future $y = (s_{\bar{t}+1}, \cdots, s_T)$ up to time T. Modern forecasting models are largely built with encoder-decoder neural networks, where the encoder $f(\cdot)$ extracts a compact representation h_i with respect to agent i and the decoder $g(\cdot)$ rolls out the predicted trajectory.

The training data $\mathcal{D}$ is often collected from a set of K environments $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$. Previous work typically merges them into a large dataset and assumes the mixture representative of the unseen test environment $\tilde{e}$. Under this assumption, the model is trained to minimize the empirical risk:

$$\mathcal{R}(f, g) := \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell(g(f(x)), y). \tag{1}$$

Unfortunately, the i.i.d. assumption does not always hold in practice. In fact, recent work [18] has shown that the test environment can be significantly different from the training ones in the widely used ETH-UCY benchmark. We will next introduce a causal formalism of motion forecasting that allows us to formulate this challenge and design solutions to address it.

Causal formalism. Motion behaviors are essentially dynamic processes governed by latent variables, such as physical laws, traffic rules and social norms. To build accurate predictive models, the conventional learning paradigm typically aims to discover these latent variables and model their correlations with the observed future states. However, the learned correlations may vary across environments and thus fail to generalize at test time.

To address this fundamental limitation, we introduce a new formalism of motion forecasting through the lens of causality. As shown in Figure 2, we explicitly categorize the latent variables into three groups:

- Invariant variables: physical laws that are universal to everyone at any place;
- Hidden confounders: motion styles that may vary from site to site in a local and sparse manner;
- Spurious features: other latent variables, such as level of noises, that are not direct causes of future motion.

In principle, neither the second group nor the third have stable correlations with the target future motion. Yet, these two groups may lead to distinctive effects on forecasting models in the presence of environmental changes. The relations with spurious features can become largely different and even completely reversed in new domains, resulting in catastrophic errors at test time. In contrast, the change of motion styles is often more restricted. While failing to capture the correct style yields inaccurate predictions, it should still output plausible solutions subject to physical laws. We will next describe two algorithmic components that treat spurious features and hidden confounders differently, in order to promote the robustness and reusability of the learned motion representations.

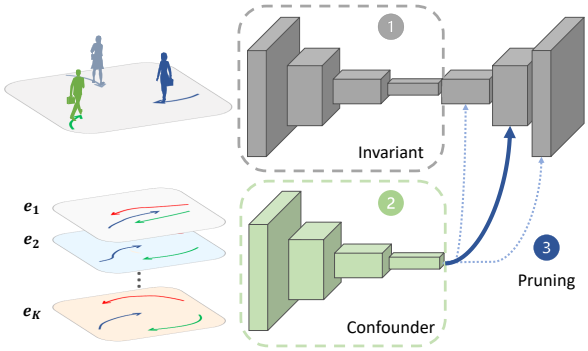

Figure 2: Our modular forecasting model contains two separate encoders for universal laws and style confounders respectively. The model is built in three steps: (i) learn an invariant predictor equally optimal in all training environments, (ii) attach an style encoder to further improve accuracy in each specific environment, (iii) prune the dense connections in the normalization layers of the decoder to approximate a sparse causal graph.

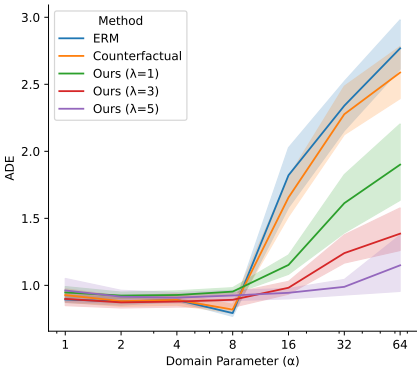

Figure 3: Comparison of different methods on the modified ETH-UCY dataset with spurious correlations. Our IRM-based approach clearly outperforms the conventional ERM and the counterfactual approach [18] in the out-of-distribution regime $\alpha \in (8, 64]$, while being on par within the training domains.

Invariant principle. By definition, invariant features have identical joint distributions with the future motion across environments, whereas the non-invariant ones are the opposite. This distinction can be converted to a necessary condition for domain invariant predictor: $g \circ f$ is equally optimal in every environment [19]. However, the exact form of this learning principle lands in a bi-level optimization, which is difficult to solve in practice. Recent work [20] proposed to relax it to a gradient norm penalty over the empirical risk $\mathcal{R}^e$ in each training environment:

$$\min_{f,g} \quad \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\mathcal{R}^e(f,g) + \lambda \|\nabla_g \mathcal{R}^e(f,g)\|_2^2 \right]. \tag{2}$$

Intuitively, the invariant risk minimization (IRM) objective prevents the forecasting model from learning an *average* effect of spurious features on future trajectories and enforces the model to solve it the hard way by seeking universal mechanisms behind motion behaviors. We show in Appendix B.1 that this technique can greatly improve the robustness of the forecasting model against the domain shifts to spurious correlations.

However, the strength of suppressing spurious features comes with a clear drawback, *i.e.*, the learned representation tends to drop the motion style as well if they change across environments. This may cause inaccurate predictions in both training and test environments, as shown in Appendix B.2. To cope with this issue, we next introduce a modular architecture that allows the model to properly structure the knowledge and strategically adapt from one style to another.

Modular network. Despite a detailed difference in architecture design, most existing forecasting models adopt dense connections at their core. On one hand, this major design choice is powerful given sufficient training data; on the other hand, the built models are often highly over-parametrized, establishing a considerable amount of unnecessary edges. Ideally, a forecasting model could preserve a clear structure of the learned knowledge, separate the impacts of physical laws and motion styles on the future motion, and approximate the sparse causal graph in Figure 2.

To achieve this goal, we devise a modular network that consists of two encoders and one decoder. The first encoder is trained to capture domain-invariant features, as described above. Subsequently, we introduce a second encoder responsible for the motion styles varying across domains. Given a small set of observations from a particular domain, the role of the style encoder is to infer the hidden confounders that influence both the history x and the future trajectory y. Unlike the conventional concatenation, we combine the two groups of causal features through normalization layers of the decoder [27], where the style features gradually modulate the prediction outputs. Under the sparse causal shift hypothesis [11], we progressively prune the connections of smaller weights during training [22]. This sparsity-driven design and training allows the model to learn a close approximation of the semantic-level causal graph, more strategically adapt the style encoder in the low data regime.

References

[1] Dirk Helbing and Peter Molnar. Social Force Model for Pedestrian Dynamics. *Physics Review E*, May 1998.

[2] Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-Body Collision Avoidance. In Cédric Pradalier, Roland Siegwart, and Gerhard Hirzinger, editors, *Robotics Research*, Springer Tracts in Advanced Robotics, pages 3–19, Berlin, Heidelberg, 2011. Springer.

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio 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. ISSN: 1063-6919.

[4] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 683–700, Cham, 2020. Springer International Publishing.

[5] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, June 2011. ISSN: 1938-7228.

[6] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social NCE: Contrastive Learning of Socially-aware Motion Representations. *arXiv:2012.11717 [cs]*, August 2021. arXiv: 2012.11717.

[7] Todor Davchev, Michael Burke, and Subramanian Ramamoorthy. Learning Structured Representations of Spatial and Interactive Dynamics for Trajectory Prediction in Crowded Scenes. *IEEE Robotics and Automation Letters*, 6(2):707–714, April 2021. Conference Name: IEEE Robotics and Automation Letters.

[8] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, November 2020. ISSN: 2640-3498.

[9] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018.

[10] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA, November 2017.

[11] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards Causal Representation Learning. *arXiv:2102.11107 [cs]*, February 2021. arXiv: 2102.11107.

[12] Yixin Wang and Michael I. Jordan. Desiderata for Representation Learning: A Causal Perspective. *arXiv:2109.03795 [cs, stat]*, September 2021. arXiv: 2109.03795.

[13] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *arXiv:2109.00725 [cs]*, September 2021. arXiv: 2109.00725.

[14] Anne Schlottmann, Deborah Allen, Carina Linderoth, and Sarah Hesketh. Perceptual Causality in Children. *Child Development*, 73(6):1656–1677, 2002.

[15] Anna Waismeyer and Andrew N. Meltzoff. Learning to make things happen: Infants' observational learning of social and physical causal events. *Journal of Experimental Child Psychology*, 162:58–71, October 2017.

[16] Stefano Pellegrini, Andreas Ess, and Luc 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.

[17] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by Example. *Computer Graphics Forum*, 26(3):655–664, 2007.

[18] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human Trajectory Prediction via Counterfactual Analysis. *arXiv:2107.14202 [cs]*, July 2021. arXiv: 2107.14202.

[19] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12167.

[20] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, March 2020. arXiv: 1907.02893.

[21] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. *arXiv:2106.06607 [cs, stat]*, June 2021. arXiv: 2106.06607.

[22] T. Obata and M. Hagiwara. Neural cognitive maps (NCMs). In *Computational Cybernetics and Simulation 1997 IEEE International Conference on Systems, Man, and Cybernetics*, volume 4, pages 3337–3342 vol.4, October 1997. ISSN: 1062-922X.

[23] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning Independent Causal Mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, July 2018. ISSN: 2640-3498.

[24] Michel Besserve, Rémy Sun, Dominik Janzing, and Bernhard Schölkopf. A theory of independent mechanisms for extrapolation in generative models. *arXiv:2004.00184 [cs, stat]*, March 2020. arXiv: 2004.00184.

[25] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent Independent Mechanisms. In *International Conference on Learning Representations*, September 2020.

[26] Kanika Madan, Nan Rosemary Ke, Anirudh Goyal, Bernhard Schölkopf, and Yoshua Bengio. Fast and Slow Learning of Recurrent Independent Mechanisms. *arXiv:2105.08710 [cs]*, May 2021. arXiv: 2105.08710.

[27] Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, October 2017. ISSN: 2380-7504.

[28] 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. ISSN: 2575-7075.

[29] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi 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.

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

[31] Deyao Zhu, Mohamed Zahran, Li Erran Li, and Mohamed Elhoseiny. Motion Forecasting with Unlikelihood Training in Continuous Space. June 2021.

[32] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-Robot Interaction: Crowd-Aware Robot Navigation With Attention-Based Deep Reinforcement Learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6015–6022, May 2019. ISSN: 2577-087X.

A Related Work

Modern motion forecasting models [3, 4, 28–30] are largely built with neural networks and trained with the maximum likelihood principle. Despite strong performance for short-range predictions within the training domain, they often struggle to generalize under covariate shift. Recently, a couple of papers proposed to promote their robustness by means of negative data augmentation [6, 31]. However, designing negative examples of high-dimension, *e.g.*, long sequences, can be difficult in practice. Our work explores a causality-inspired alternative that does not require hand-engineered interventions over training data and is hence more theoretically elegant and algorithmically generic.

Closedly related to ours, another recent work [18] attempts to mitigate dataset biases in motion forecasting problems through counterfactual analysis. Our method differs from theirs in three aspects: (i) their approach needs to explicitly estimate dataset biases, whereas our method aims to directly suppress the biased features; (ii) their approach inherits the merge-and-shuffle convention, which destroys the critical information for bias estimation; in contrast, we keep each subset separately and exploit varied correlations across environments; (iii) their adopted approach – counterfactual – is known to be the most difficult in Pearl's ladder of causation [9]; conversely, our method formulates spurious correlations from the intervention perspective, which is more viable to tackle in practice.

B Experiments

We evaluate our proposed method on human motion forecasting under distributional shifts of two types of variables, namely spurious features or style confounders. In the considered forecasting task, a model reads the past 8 time steps (3.2 seconds) of human trajectories in the scene and predicts their future movements in the following 12 (4.8 seconds) time steps. Throughout our experiments, we use the STGAT [29], a state-of-the-art trajectory forecasting model, as the base model. It contains two LSTMs and one graph attention network (GAT) to encode the historical trajectory and social interaction clues, and an LSTM decoder for prediction.

We compare the forecasting model trained by our IRM-based method with those trained by previous baselines:

- *Empirical Risk Minimization* (ERM): the conventional learning method that minimizes the average prediction error on all training samples (Eq. 1);
- *Counterfactual Analysis* [18]: a recent causality-inspired trajectory forecasting method that estimates and subtracts biased features through counterfactual interventions.

Following prior works [3, 4, 28], we evaluate forecasting models on two metrics:

- *Average Displacement Error* (ADE): the average Euclidean distance between the predicted output and the ground truth over all predicted time steps.
- *Final Displacement Error* (FDE): the Euclidean distance between the predicted final destination and the true final destination at end of the prediction horizon T.

We evaluate each method over five experiments with different random seeds.

B.1 Controlled Spurious Features

Setup We first validate the effectiveness of our proposed IRM-based method on a modified ETH-UCY datasets with a controlled spurious feature. The original ETH-UCY dataset contains five subsets collected at different locations. Recent work [18] has shown that these subsets present intrinsic differences. However, it is not trivial to pinpoint the detailed biases in each environment. To clearly examine the robustness of a motion forecasting model against non-causal biases, we modified the ETH-UCY dataset by introducing a third confidence variable. Apparently, the level of confidence about the past observation is not a direct cause of future motion, yet it may bring information about the local curvature of the trajectory. We simulate this phenomenon and assume the confidence score c is a linear function of the local curvature in a short temporal window:

$$c_t^i := \alpha \cdot [(\dot{x}_{t+\delta t}^i - \dot{x}_t^i)^2 + (\dot{y}_{t+\delta t}^i - \dot{y}_t^i)^2 + 1], \qquad \forall t = 1, \cdots, \bar{t} \tag{3}$$

where $\dot{x}_t^i = x_{t+1}^i - x_t^i$ and $\dot{y}_t^i = y_{t+1}^i - y_t^i$ reflect the velocity of the agent, $\delta t = 8$ is the length of the temporal window, and α is an environment-dependent coefficient. We train the model in four environments ('hotel','univ','zara1' and 'zara2') with $\alpha \in \{1, 2, 4, 8\}$ and test it on the last one ('eth') with $\alpha \in \{1, 2, 4, 8, 16, 32, 64\}$.

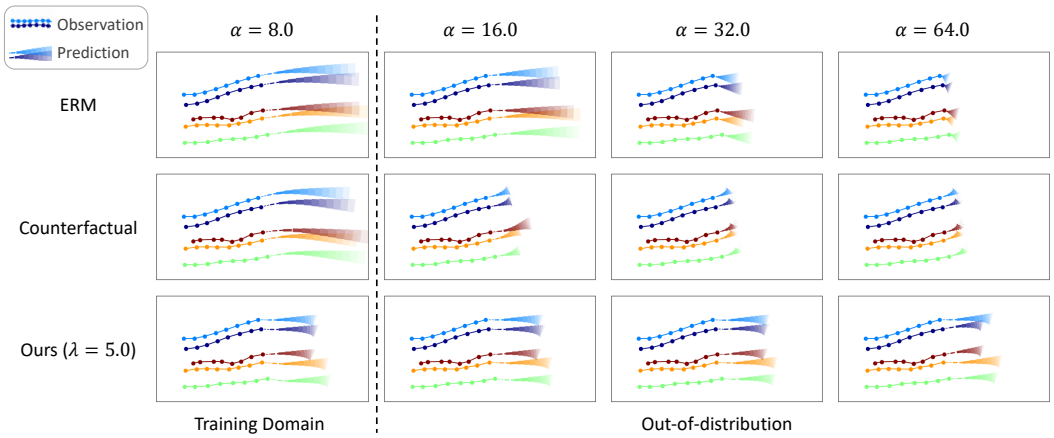

Figure 4: Visualization of the predicted trajectories from different learning methods in the presence of spurious features. The future trajectories predicted by the two baselines abruptly slow down, when the domain parameter α at test time is larger than that in the training domains. In comparison, our IRM-based approach results in much more robust solutions, even under substantial changes of the spurious feature (*e.g.*, $\alpha = 64.0$).

Method	Train	Val	Test
ERM	$\mathbf{0.525 \pm 0.034}$	$\mathbf{0.525 \pm 0.034}$	0.607 ± 0.007
Counterfactual	0.531 ± 0.012	0.531 ± 0.012	$\mathbf{0.587 \pm 0.008}$
Ours ($\lambda = 10^{-3}$)	0.560 ± 0.013	0.559 ± 0.012	0.602 ± 0.005
Ours ($\lambda = 10^{-2}$)	0.563 ± 0.019	0.563 ± 0.018	0.623 ± 0.020
Ours ($\lambda = 0.10$)	0.612 ± 0.017	0.612 ± 0.018	0.623 ± 0.006
Ours ($\lambda = 1.00$)	0.642 ± 0.007	0.641 ± 0.007	0.661 ± 0.003

Table 1: Quantitative comparison of different methods on the synthetic dataset with style shifts. The prediction accuracy is measured by FDE and evaluated over 5 seeds. The baseline methods suffer from clear performance drops in the test domains due to its 'average' style of the training domains. The IRM-based approach is not able to address the change of style confounder neither, as it tends to ignore the domain-specific style information. This result suggests the importance of designing adaptable forecasting models in the motion context.

Results Figure 3 shows the prediction accuracy on the test sets resulted from different learning methods. The two baseline methods degrades drastically under large shifts of the domain parameter. In comparison, the forecasting models trained by our IRM-based method are much less sensitive to domain shifts. The qualitative results in Figure 4 also confirm this observation.

B.2 Synthetic Style Confounders

Setup We further evaluate forecasting models trained by our method under style shifts. The motion styles of existing real-world data are often largely unknown. We, therefore, create some synthetic trajectories using ORCA [2], a popular multi-agent simulator, in the circle-crossing scenarios [32] with varied style parameters. Specifically, we consider four training styles, where the simulated agents keep different minimum separation distances from each other, *i.e.*, $\{0.15, 0.20, 0.25, 0.30\}$ meters. For each training domain, we generate 10,000 trajectories for training and 3,000 trajectories for validation. We test the model on 10,000 trajectories in two new environments with the minimum separation distance of 0.35 or 0.40 to examine its robustness in the out-of-distribution regime.

Results Table 1 reports the results of motion forecasting models on the training, validation, and test sets using different methods. Given style changes, the IRM-based method does not yield any improvements over the baseline methods. In fact, it leads to worse prediction accuracies in both training and test domains, as it enforces the model to ignore the domain-specific style confounder, which is however essential for future motion. This result reveals the limit of IRM in the motion context and confirms the importance of building adaptable forecasting models.