A Physics-Informed In-Context Learning Framework for Online Interaction Prediction in Robotic Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract:
The estimation of interaction forces is a key enabler for safe and adaptive physical interactions, especially for robotics. Yet, it remains a challenging problem under distributional shifts (*e.g.*, time-varying and uncertain dynamics). In this work, a novel physics-informed in-context learning framework for real-time force prediction in robotic manipulation is introduced. Our approach extends transformer-based meta learning with physically grounded inductive biases, including learnable physics parameters, physics-aware embeddings, and regularization via physics-based loss functions. The model is trained on a real-world interaction dataset collected from a robotic arm executing chirp-like (*i.e.*, multi-frequency) trajectories against compliant surfaces. We evaluate generalization to out-of-distribution scenarios involving unseen materials and excitation signals, showing that our method significantly outperforms both standard Transformer-based models and model-based baselines such as the Extended Kalman Filter in the most challenging settings. Furthermore, we validate the model's online predictive capabilities in closed-loop deployments, showcasing its applicability in downstream control tasks.

Keywords: In-context learning, physics-informed neural networks, robot interaction force prediction

1 Introduction

Accurate estimation and prediction of interaction forces are critical in modern robotic applications [1]. In human–robot collaboration (HRC), for instance, precise knowledge of the forces exchanged during physical contact is essential to ensure safety, enhance responsiveness, and enable seamless and intuitive cooperation between human and robotic partners [2, 3, 4]. Several studies have demonstrated that even slight inaccuracies in force estimation can lead to misinterpretations of human intent and cause unstable robot behaviors, posing significant safety risks [5, 6]. Traditional force estimation methods, which rely on sensor fusion and model-based observers, perform well in structured environments [7]; however, they often struggle when operating in dynamic ones, as shown for instance in [8], or unstructured settings where the interaction forces vary rapidly and unpredictably.

To overcome these limitations, meta learning has emerged as a compelling paradigm for building models that rapidly adapt to new tasks with minimal data. Meta learning has recently gained renewed attention due to its ability to generalize across tasks via shared inductive biases [9]. Popular frameworks include Model-Agnostic Meta Learning (MAML) [10] and in-context learning (ICL) [11], where models leverage contextual cues from input sequences to adapt on-the-fly. A key advantage of ICL is its zero-shot generalization capability: the model is pre-trained on a diverse set of tasks and can generalize to new, unseen tasks during inference without requiring fine-tuning. This ability, popularized by large language models [12], offers a compelling path toward flexible, data-efficient adaptation. However, one of the main limitations of ICL lies in its sensitivity to out-of-distribution

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

(OOD) generalization. Despite its impressive performance on in-distribution tasks, current research shows that in-context learners can struggle when deployed in environments that differ significantly from their training distribution [13, 14]. Improving robustness and generalization in OOD scenarios remains an open and active area of research.

To this end, this project proposes a novel physics-informed in-context meta learner, hereafter referred to as the meta model. This approach builds upon the ICL framework introduced by [15], extending it to integrate domain knowledge from physics directly into the model's architecture and training objective. The meta model is trained to predict interaction forces using real-world datasets, each capturing the behavior of a robotic arm interacting with a specific material. During inference, it is deployed in a zero-shot fashion on a completely different unseen material, scenarios that are explicitly OOD with respect to training data. To guide generalization, the architecture introduced in this paper incorporates physics-based loss functions, physics-aware input embeddings, and physics-based positional encodings. Attention mechanisms are also adapted to emphasize physically meaningful features. The meta model predicts not only the current interaction force but also its future evolution, an essential capability for anticipating dynamics, *e.g.*, in HRC and advanced manufacturing [16]. The predicted interaction forces can be leveraged by a controller, such as a model predictive controller [17], enabling robots to more effectively anticipate the interaction. However, the controller design is out of scope for this paper. With respect to [15], a major contribution consists in the regularization of the in-context learner by embedding physical structure, improving its robustness and generalization in challenging OOD settings, while retaining its ability to adapt without fine-tuning.

2 Related Work

Interaction-Force Estimation: Model-Based vs Data-Driven approaches. Estimating interaction forces in robot manipulators can be divided into two paradigms: model-based observers, which leverage analytical dynamic models (*e.g.*, Euler–Lagrange formulations [18], compliance models [19, 20]), and data-driven (or hybrid) methods that learn force mappings from sensor data (e.g., neural networks [21, 22], Gaussian processes [23]). Model-based observers offer interpretability and computational efficiency but degrade under unmodeled friction, payload changes, and stiffness uncertainty [24]. Data-driven models can capture complex contact phenomena and adapt to new conditions, but rely on large datasets and may generalize poorly out of distribution. Additionally, physics-informed machine learning approaches have been developed to combine the advantages of model-based and data-driven methods [25, 26]. Most existing studies focus on estimating the current contact forces [27], whereas anticipative and compliant control can benefit from predicting future force trajectories. Traditional sequence models (RNNs, LSTMs) suffer from sequential processing bottlenecks and limited long-horizon performance [28], motivating transformer architectures with self-attention to capture long-range dependencies more effectively [29].

Transformers for Contact Force Prediction. To address the need for accurate force prediction, transformer models have recently been investigated for dynamical modeling. Geneva *et al.* [30] demonstrated that transformers can learn dynamics governed by partial differential equations, outperforming RNN and CNN baselines. Building on this, Fusco *et al.* [27] proposed a transformer-based observer for human–robot interaction. Domínguez *et al.* [31] suggested replacing LSTMs with transformers for force prediction and demonstrated preliminary gains in accuracy. In follow-up work, they used Vision Transformer and Swin Transformer backbones to fuse complex sensory data, jointly predicting human-applied forces and future motion states [32]. Despite these advances, existing transformer-based methods for contact force estimation typically rely on rich multimodal inputs (*e.g.*, mocap data, LiDAR scans, vision) and introduce significant sensor and computational complexity. Additionally, they have not been tested in OOD scenarios, especially in dynamic and unpredictable real-world environments.

Transformer-Based In-Context Learning. The concept of ICL was first introduced in the foundational work by [11], where transformer-based models demonstrated the ability to perform tasks using only inference-time input-output examples, without requiring parameter updates. Since then, various studies have analyzed its theoretical properties and limitations [33, 34], while others have expanded its application to domains such as *robotics* [35, 36], *control* [37], *system identification* [38], and *vision* [39]. Despite its flexibility, a central challenge for ICL remains its limited generalization to OOD contexts [13, 14]. The physics-informed variant of ICL introduced in this paper is designed to improve OOD robustness by integrating physical principles into the ICL framework.

3 Method

Transformer-based architectures have been widely adopted in the ICL framework, with applications ranging from large-scale language modeling [40] to robotics control and decision-making tasks [36]. This section first reviews the standard, purely data-driven ICL approach using transformers, which has been adopted as a baseline in this work, and then introduces the proposed physics-informed extension designed to enhance performance, particularly in OOD scenarios.

3.1 Data-Driven Approach: Standard Transformer

The architecture employed in this paper is based on the one proposed by [15], which itself is an adaptation of the original transformer architecture [29], modified to process real-valued time-series input/output sequences instead of discrete word tokens. In this paper's implementation, the authors adopt the above-mentioned architecture with minor adjustments and apply it to the task of force prediction.

The meta model $\mathcal{M}_\phi$ operates over a set of input/output datasets $\{\mathcal{D}^{(i)} = (u^{(i)}_{1:N}, y^{(i)}_{1:N}),\ i = 1, 2, \dots\}$, where $u^{(i)}_k \in \mathbb{R}^{n_u}$ and $y^{(i)}_k \in \mathbb{R}^{n_y}$ denote input and output vectors at time step k for dataset i, generated by distinct dynamical systems that *share latent structure*. The input u has dimension n_u = 15 and contains the Cartesian positions (x, y, z), velocities $(\dot{x}, \dot{y}, \dot{z})$, accelerations $(\ddot{x}, \ddot{y}, \ddot{z})$, along with target positions (x_T, y_T, z_T) and velocities $(\dot{x}_T, \dot{y}_T, \dot{z}_T)$. The output y has dimension $n_y = 3$ and consists of the interaction forces (F_x, F_y, F_z) measured at the end-effector. Each dataset $\mathcal{D}^{(i)}$ is split into a context consisting of the first m samples and a query containing the last m samples. The prediction horizon p is selected such that $p < m$, ensuring partial overlap between the context and query segments.

For each dataset, given the full context data $(u^{(i)}_{1:m}, y^{(i)}_{1:m})$ and query inputs $u^{(i)}_{p+1:N}$, where $N = m+p$ and $p < m$, the meta model predicts the corresponding query outputs $\hat{y}^{(i)}_{p+1:N}$ as

$$\hat{y}^{(i)}_{p+1:N} = \mathcal{M}_\phi(u^{(i)}_{p+1:N}, u^{(i)}_{1:m}, y^{(i)}_{1:m}), \tag{1}$$

which are then compared to the ground truth $y^{(i)}_{m+1:N}$, during training, by minimizing the Mean Squared Error (MSE) loss J over a minibatch of b datasets:

$$J = \frac{1}{b} \sum_{i=1}^{b} \left\| y^{(i)}_{m+1:N} - \mathcal{M}_\phi(u^{(i)}_{m+1:N}, u^{(i)}_{1:m}, y^{(i)}_{1:m}) \right\|^2 . \tag{2}$$

The core idea behind the ICL framework is to enable the model to infer latent dynamics from a context window and generalize to future behavior within a query window. By training across multiple datasets governed by distinct dynamics, the meta model learns *common structure*, allowing zero-shot generalization at inference time without fine-tuning. However, ICL's performance depends heavily on the distributional similarity between training and test conditions. When deployed in OOD settings, characterized by unseen dynamics or feature shifts, generalization may degrade significantly.

To mitigate this, we propose a physics-informed extension of the transformer-based ICL model. The proposed architecture improves robustness under OOD scenarios by embedding domain-specific physical priors directly into the learning process, as detailed in the following section.

3.2 Physics Informed Approach: Regularized Transformer

Embedding physical laws into neural networks has proven effective in enhancing generalization to OOD scenarios [26], as the incorporation of governing physical equations encourages the network to adhere to known physical principles. The proposed physics-informed ICL framework merges the advantages of introducing physical laws with the impressive performance of ICL. The solution is characterized by one or more of the following modifications to the data-driven approach detailed in the previous section.

Physically Enhanced Attention Mechanisms. Traditional self-attention and cross-attention mechanisms compute attention scores based solely on learned embeddings of input data, without explicit knowledge of the physical relevance of different features. In the case at hand, not all input features contribute equally to the prediction. By introducing a physics-based bias, the attention mechanism is guided to focus on time steps and features that have a stronger physical relationship with the output forces. The bias is calculated from a concatenation of relevant physical features and then projected into the attention space. This modification helps the self-attention to assign higher attention scores to timesteps where forces, velocities, or accelerations provide meaningful information. Thanks to the mentioned bias, the decoder will focus on the encoder time steps that are most relevant for predicting future interaction forces. A similar bias is embedded into the positional encoding.

Physics-Based Embedding Layers. Transformers typically use embeddings to map raw input features into a high-dimensional space where meaningful patterns can be learned. Traditional embedding layers transform input features using learnable weights without directly considering the underlying physical principles governing the system. Hence, embedding the data in a way that reflects key physical properties would be more in line with the purpose of the developed model. Therefore, a governing physical equation is introduced into the embedding layers, in which the physical parameters, like inertia J, damping b, and stiffness k, are estimated by the transformer, and shared with the physics-based loss function.

Physics-Based Loss Function. A physics-informed loss term J_{Phys}, computed as the MSE between the model's outputs and those generated by the underlying physical model, is added to ensure compliance with the governing laws:

4 Validation and Comparisons

4.1 Dataset Creation

To gather data, a position-level chirp signal is applied to generate interaction forces between the robot and its environment, as it is a common content-rich excitation signal for dynamic identification procedures [41]. Its time-varying frequency excites the system over a wide bandwidth, enabling efficient and high-resolution extraction of frequency-domain characteristics in a single experiment [42]. The reference-motion signal is, indeed, defined as follows:

$$z^r(t) = z^0 + \delta \cos\left(2\pi h_1 \left(1 + \frac{1}{20}\cos(2\pi h_2 t)\right) t\right), \tag{3}$$

where z^0 denotes the initial position of the chirp, and δ its amplitude, both selected to guarantee contact with the environment at all times during the experiment, $h_1 = 0.9$ and $h_2 = 0.127$ are frequency-related coefficients, and t is the time. The goal was to cover interaction forces from 2 N to approximately 18 N, with most environments, when applicable, and to excite the system up to the required frequencies.

While all data are governed by the same underlying mechanical interaction principles, each dataset corresponds to a different physical environment, *e.g.*, the robot interacting with materials of varying stiffness. These variations induce significant shifts in the dynamics and force profiles due to differing material properties and human-induced variability in experimental setups. Such variability presents a major challenge to generalization.

4.2 Interaction Task Description

The chirp signal in (3) is used as the reference position input to implement a probing task. The low-level robot controller is fed with such a signal along the z-axis (*i.e.*, the axis perpendicular to the table plane), while the remaining position and orientation components are held fixed. A Cartesian impedance controller [7] tracks this reference through its commanded pose $\mathbf{x}^d$ (which includes both position and orientation references), enabling the robot to interact with its environment safely.

4.3 Materials

Robotic platform: experiments were conducted using a Franka EMIKA Panda 7-DoF manipulator, equipped with integrated joint-torque sensors. The robot was controlled via a Cartesian impedance controller operating at 1 kHz, implemented in ROS Noetic on Ubuntu 20.04.
Interaction fixtures: physical interactions were performed against a rigid table and four compliant foam sponges of varying stiffness (see Section 4.1). The end-effector was constrained to the vertical (z) axis; all other axes and orientations remained fixed throughout the experiments.
Computing platform: model training and evaluation were conducted on a laptop equipped with an Intel Core Ultra 9 processor, 32 GB RAM, and an NVIDIA RTX 4080 GPU with 12 GB of VRAM.

4.4 Training

The datasets were split into training and validation subsets, with proportions of 85% and 15%, respectively. All subsets are segmented into non-overlapping sequences, then split into context and query, each of length $m = 352$, while the transformer is trained to predict the force over a window size $p = 240$. Therefore, as explained in Section 3.1, the context and query overlap by $m - p = 112$ datapoints. Specifically, if the context sequence spans from time 0 to 352, the corresponding query sequence starts at time 240 and extends to 592. This intentional overlap allows the model to use the most recent data from the context sequence to predict the query sequence. It's worth to remark that the overlapping segments are excluded from the loss computation and all reported metrics.

Both transformer-based meta models feature $n_{\text{layers}} = 8$ layers and a hidden dimension of $d_{\text{model}} = 128$, using $n_{\text{heads}} = 4$ attention heads in both encoder and decoder modules. Training proceeds for up to 4000 epochs using the AdamW optimizer [43]. At each epoch, one dataset is selected at random, and a mini-batch of size $b = 16$ is sampled. Early stopping is employed based on performance across randomized validation subsets to mitigate overfitting. At each batch, the encoder and decoder input statistics (*i.e.*, mean μ and standard deviation δ) were computed and used to normalize the meta model's input; moreover, the forces from the encoder were also used to denormalize the predictions.

References

[1] L. Roveda, F. Vicentini, and L. M. Tosatti. Deformation-tracking impedance control in interaction with uncertain environments. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1992–1997. IEEE, 2013.

[2] M. Maccarini, F. Pura, D. Piga, L. Roveda, L. Mantovani, and F. Braghin. Preference-based optimization of a human-robot collaborative controller. *IFAC-PapersOnLine*, 55(38):7–12, 2022.

[3] D. Riedelbauch, N. Höllerich, and D. Henrich. Benchmarking teamwork of humans and cobots—an overview of metrics, strategies, and tasks. *IEEE Access*, 11:43648–43674, 2023.

[4] S. SMBPB, M. Valori, G. Legnani, and I. Fassi. Assessing safety in physical human–robot interaction in industrial settings: A systematic review of contact modelling and impact measuring methods. *Robotics*, 14(3):27, 2025.

[5] T. T. Mac, C. Copot, D. T. Tran, and R. De Keyser. Heuristic approaches in robot path planning: A survey robotics and autonomous systems. 2016.

[6] S. Haddadin and E. Croft. Physical human–robot interaction. In *Springer handbook of robotics*, pages 1835–1874. Springer, 2016.

[7] L. Roveda, D. Riva, G. Bucca, and D. Piga. Sensorless optimal switching impact/force controller. *IEEE Access*, 9:158167–158184, 2021.

[8] Y. Wang, D. Held, and Z. Erickson. Visual haptic reasoning: Estimating contact forces by observing deformable object interactions. *IEEE Robotics and Automation Letters*, 7(4):11426–11433, 2022.

[9] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[10] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[13] K. Ahuja and D. Lopez-Paz. A closer look at in-context learning under distribution shifts. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*. PMLR, 2023.

[14] Q. Wang, Y. Wang, X. Ying, and Y. Wang. Can in-context learning really generalize to out-of-distribution tasks? In *The Thirteenth International Conference on Learning Representations*. PMLR, 2025.

[15] M. Forgione, F. Pura, and D. Piga. From system models to class models: An in-context learning paradigm. *IEEE Control Systems Letters*, 7:3513–3518, 2023.

[16] L. Roveda, N. Pedrocchi, M. Beschi, and L. M. Tosatti. High-accuracy robotized industrial assembly task control schema with force overshoots avoidance. *Control Engineering Practice*, 71:142–153, 2018.

[17] L. Roveda, A. Testa, A. A. Shahid, F. Braghin, and D. Piga. Q-learning-based model predictive variable impedance control for physical human-robot collaboration. *Artificial Intelligence*, 312:103771, 2022.

[18] M. W. Spong. On the robust control of robot manipulators. *IEEE Transactions on automatic control*, 37(11):1782–1786, 1992.

[19] L. Roveda, A. Bussolan, F. Braghin, and D. Piga. 6d virtual sensor for wrench estimation in robotized interaction tasks exploiting extended kalman filter. *Machines*, 8(4):67, 2020.

[20] L. Roveda, L. Mantovani, M. Maccarini, F. Braghin, and D. Piga. Optimal physical human–robot collaborative controller with user-centric tuning. *Control Engineering Practice*, 139: 105621, 2023.

[21] J. Hu and R. Xiong. Contact force estimation for robot manipulator using semiparametric model and disturbance kalman filter. *IEEE Transactions on Industrial Electronics*, 65(4):3365–3375, 2017.

[22] W. Deng, F. Ardiani, K. T. Nguyen, M. Benoussaad, and K. Medjaher. Physics informed machine learning model for inverse dynamics in robotic manipulators. *Applied Soft Computing*, 163:111877, 2024.

[23] Y. Wei, S. Lyu, W. Li, X. Yu, Z. Wang, and L. Guo. Contact force estimation of robot manipulators with imperfect dynamic model: on gaussian process adaptive disturbance kalman filter. *IEEE Transactions on Automation Science and Engineering*, 2023.

[24] Y. Han, J. Wu, C. Liu, and Z. Xiong. An iterative approach for accurate dynamic model identification of industrial robots. *IEEE Transactions on Robotics*, 36(5):1577–1594, 2020.

[25] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[26] J. Liu, P. Borja, and C. Della Santina. Physics-informed neural networks to model and control robots: A theoretical and experimental investigation. *Advanced Intelligent Systems*, 6(5): 2300385, 2024.

[27] A. Fusco, V. Modugno, D. Kanoulas, A. Rizzo, and M. Cognetti. Transformer-based prediction of human motions and contact forces for physical human-robot interaction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3161–3167. IEEE, 2024.

[28] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] N. Geneva and N. Zabaras. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, 2022.

[31] J. E. Domínguez-Vidal and A. Sanfeliu. Improving human-robot interaction effectiveness in human-robot collaborative object transportation using force prediction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7839–7845. IEEE, 2023.

[32] J. E. Domínguez-Vidal and A. Sanfeliu. Exploring transformers and visual transformers for force prediction in human-robot collaborative transportation tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3191–3197. IEEE, 2024.

[33] V. Herrmann, R. Csordás, and J. Schmidhuber. Measuring in-context computation complexity via hidden state prediction. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

[34] L. Kirsch, J. Harrison, J. Sohl-Dickstein, and L. Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

[35] M. B. Bazzi, A. A. Shahid, C. Agia, J. Alora, M. Forgione, D. Piga, F. Braghin, M. Pavone, and L. Roveda. Robomorph: In-context meta-learning for robot dynamics modeling. *arXiv preprint arXiv:2409.11815*, 2024.

[36] J. Y. Zhu, C. G. Cano, D. V. Bermudez, and M. Drozdzal. InCoRo: In-Context Learning for Robotics Control with Feedback Loops. *arXiv preprint arXiv:2402.05188*, 2024.

[37] R. Busetto, V. Breschi, M. Forgione, D. Piga, and S. Formentin. One controller to rule them all. *arXiv preprint arXiv:2411.06482*, 2024.

[38] M. Rufolo, D. Piga, G. Maroni, and M. Forgione. Enhanced transformer architecture for in-context learning of dynamical systems. *arXiv preprint arXiv:2410.03291*, 2024.

[39] Y. Zhou, X. Li, Q. Wang, and J. Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024.

[40] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*, 2023.

[41] L. Roveda, G. Pallucca, N. Pedrocchi, F. Braghin, and L. M. Tosatti. Iterative learning procedure with reinforcement for high-accuracy force tracking in robotized tasks. *IEEE Transactions on Industrial Informatics*, 14(4):1753–1763, 2017.

[42] O. Nelles. *Nonlinear dynamic system identification*. Springer, 2020.

[43] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.