Standards-Compliant Active-Inference Framework for Safe Human-Robot Collaboration

Alessio Caneschi[12*], Kevin Alcedo[34*], Pedro U. Lima[34], Matteo Bottin[1], Giulio Rosati[1]
[1]Department of Industrial Engineering, University of Padova
[2]Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS)
[3]Institute for Systems and Robotics / LARSyS
[4]Lisbon ELLIS Unit, Instituto Superior Técnico, Universidade de Lisboa
[*]Equal contribution: alessio.caneschi@studenti.unipd.it, kevin.alcedo@tecnico.ulisboa.pt

***Abstract*—Recent advances in manufacturing have positioned human-robot collaboration as a key enabler of flexible production, placing increasing demand on automation systems capable of anticipating human motion rather than merely reacting to it. Purely reactive safety can be overly conservative, while fixed motion forecasting can be brittle under missing observations or ambiguous human subgoals. We propose a probabilistic short-horizon active-inference framework over robot motion primitives, with epistemic value restricted to uncertainty over the human's current subgoal in a known ordered task. A learned human-motion model conditions the forecast, while a safety shield bounds predicted collision energy in shared pick-and-place. Individual components have been implemented and evaluated; ongoing work integrates them toward standards-compliant collision mitigation.**

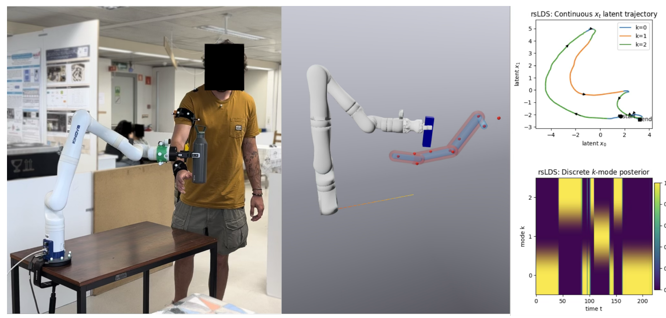

Fig. 1. Experimental setup and human-motion modeling results. Left: Kinova Gen3 Lite robot and motion-capture markers. Center: Drake simulation. Right: task-aware model results from motion-capture data, showing learned switching latent dynamics (top) and rsLDS mode-transition structure (bottom).

I. Introduction & Background

Collaborative industrial robot deployments are governed by standards such as ISO 10218-1/2:2025, which incorporate collaborative-application guidance from ISO/TS 15066. These standards provide essential safety requirements, but practical deployments often rely on conservative assumptions that can reduce task efficiency when human motion, task context, and contact conditions vary. This motivates robot controllers that are not only standards-aware, but also context-aware and uncertainty-aware.

Towards this vision, our work aims to integrate recent work on (1) human-robot collision risk assessment (Section II-A), (2) task-aware probabilistic hybrid dynamical human-motion modeling (Section II-B), and (3) active-inference robot planning under uncertainty (Section II-C). We envision a deployment that infers task-aware human dynamics and intent, and adapts robot plans to actively reduce uncertainty while maintaining safe, standards-compliant collision mitigation.

This work was supported by LARSyS FCT funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020), the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI. This research was also funded by the Italian PNRR (Ministerial Decree No. 351 of 9 April 2022, grant: 1546). The results of this research were fully obtained within the framework of the Italian National PhD Program in Robotics and Intelligent Machines.

II. Approach

A. Safe & Standards-Compliant Collision Modeling

We consider collaborative tasks where incidental human-robot contact is permissible when it remains within the limits prescribed by relevant standards and regulations, specifically ISO 10218-1/2:2025, which also includes ISO/TS 15066. Strictly contact-free policies can lead to unnecessary delays, stop-work time, or awkward interactions. We aim to show that dynamically controlling robot energy, conditioned on task-aware human-motion forecasts, improves task performance relative to policies that avoid all contact and ignore human intent.

Standards-compliant[1] collisions satisfy:

$$E = \frac{F^2}{2k_{\text{eff}}} = \frac{1}{2}\mu v_{\text{rel}}^2, \tag{1}$$

where for this work $k_{\text{eff}} = (k_{\text{body}}^{-1} + k_{\text{damp}}^{-1})^{-1}$ combines body-region and end-effector stiffness, and the reduced mass is:

$$\mu = \left[\frac{1}{m_H} + \frac{1}{m_R(\mathbf{q}, \hat{n})}\right]^{-1}, \tag{2}$$

with m_H from ISO 10218-2:2025 tables and $m_R(\mathbf{q}, \hat{n}) = \left(\hat{n}^\top [JM(\mathbf{q})^{-1}J^\top]^{-1}\hat{n}\right)^{-1}$ the posture-dependent robot effective mass [1], superseding the normative approximation (in which $m_R = M/2 + m_L$, with m_L the payload mass).

[1]Defined in ISO/TS 15066, ISO/PAS 5672, and ISO 10218-2:2025.

At deployment, the robot constrains v_R and configuration-dependent m_R to satisfy the applicable collision-energy limit, leveraging recent human-robot impact-force models [1], [2].

B. Task-aware Human-Motion Modeling

Inspired by structured world models from [3], we model hybrid human motion with a recurrent switching linear dynamical system (rsLDS) [4] with continuous state x_t, discrete mode $z_t \in \{0, \ldots, K-1\}$, and marker observation y_t:

$$\begin{aligned} z_t \mid z_{t-1} = j, x_{t-1} &\sim \mathrm{Cat}(\mathrm{softmax}(W_j x_{t-1} + c_j)), \\ x_t \mid z_t = k, x_{t-1} &\sim \mathcal{N}(A_k x_{t-1} + b_k, Q_k), \\ y_t \mid x_t &\sim \mathcal{N}(C x_t + d, R). \end{aligned} \tag{3}$$

Where $\mathrm{Cat}(\cdot)$ is categorical; softmax is over the K modes; and (W_j, c_j), (A_k, b_k, Q_k), and (C, d, R) are learned transition, dynamics, and emission parameters. Each observation has a visibility mask r_t; missing coordinates are omitted from the emission likelihood, $\log p(y_t \mid x_t, r_t) = \log p(y_{t,\mathrm{obs}} \mid x_t)$, essential for motion-capture sequences with rigid-body dropout.

A task-level Hidden Markov Model (HMM) tracks progress through a known ordered task script $\mathcal{T}_e = (\gamma_{e,1}, \ldots, \gamma_{e,L_e})$, where each $\gamma_{e,\ell}$ denotes a semantic subgoal, such as grasping an object or placing it in a bin for episode e. We treat task-level human intent as the current subgoal index $g_t \in \{1, \ldots, L_e\}$. Target-relative features are computed for each candidate subgoal, yielding the posterior $q_t^g(\ell) = p(g_t = \ell \mid y_{0:t}, r_{0:t}, \mathcal{T}_e)$. At runtime, a Rao-Blackwellized particle filter [5] samples rsLDS mode paths while marginalizing the linear-Gaussian state with Kalman updates. Forecasting is a prediction-only rollout from the filtered particle bank, providing causal human-motion estimates for closed-loop robot planning.

C. Robot Active Inference

Active inference has recently shown promise as a principled method for simultaneous perception and decision-making under uncertainty in robotic applications [6]. Relative to standard POMDPs and control-as-inference formulations, it emphasizes expected free energy: a trade-off between preferred outcomes and expected information gain about hidden states. We use this view to select policies over a small library of robot motion primitives, e.g., retreat, pregrasp, pick, move-to-bin, and place. Candidate policies $\pi^R_{t:t+N}$ over horizon N are rolled out using robot kinematics, workspace updates, and human-motion forecasts. The planner minimizes the expected-free-energy objective:

$$\begin{aligned} G(\pi^R) &= J(\pi^R) - \beta \mathcal{I}_g(\pi^R), \\ J(\pi^R) &= \mathbb{E}[C_{\mathrm{task}} + \lambda_s C_{\mathrm{safety}} + \lambda_d C_{\mathrm{delay}} \mid S_t, \pi^R], \end{aligned} \tag{4}$$

where $C_{\cdot}$ denote costs, S_t contains the filtered human, task, workspace, and robot states, and $\beta \geq 0$ weights information-seeking behavior relative to the task-directed cost. The information-gain term is restricted to uncertainty over the human's current subgoal:

$$\mathcal{I}_g(\pi^R) = H(q_t^g) - \mathbb{E}_{\tilde{y} \sim p(\cdot \mid S_t, \pi^R)} \left[H(q_{t+N}^g) \right], \tag{5}$$

where $H(\cdot)$ is entropy and $\tilde{y}$ is the future observation sequence predicted under π^R. Thus, the robot values policies expected to clarify human intent while keeping task, delay, and safety costs low. A safety shield rejects or modifies unsafe primitive execution using the conditions defined in Section II-A. The active inference objective improves nominal behavior under uncertainty, while the shield enforces safety constraints during execution.

III. Preliminary Results & Future Work

The core framework components have been evaluated independently in simulation and on hardware. The setup uses a Kinova Gen3 Lite, wrist-mounted RGB-D camera, human-arm motion-capture markers, and an external camera for workspace-state estimation. The framework leverages Drake for model-based robot planning and control [7], and existing structured Bayesian inference tools [8], [9]. Figure 1 shows the hardware setup, the simulation counterpart, and representative results of the rsLDS/HMM/RBPF for human-arm motion in a pick-and-place task.

Future work will integrate the safety shield, human-intent inference, motion forecasting, and active-inference planner into a closed-loop shared pick-and-place system. We will compare our integrated solution against standards-compliant collision mitigation policies, evaluating whether active-inference planning improves task efficiency while preserving standards' safety limits. Metrics include task success, completion time, robot idle time, minimum human-robot distance, shield interventions, and forecast/subgoal prediction error under observation dropout, task context, and human-subgoal uncertainty.

References

[1] A. Caneschi, M. Bottin, A. Doria, A. Cesaro, and G. Rosati, "Lumped parameters robot models to study impact dynamics," in *New Trends in Mechanism and Machine Science*. Springer Nature Switzerland, 2024, pp. 253–260.

[2] A. Caneschi, M. Bottin, and G. Rosati, "Modeling human–robot impact dynamics in collaborative applications," *Actuators*, vol. 15, no. 3, 2026.

[3] L. Da Costa, S. Namjoshi, M. A. Ansari, and B. Schölkopf, "Natural Building Blocks for Structured World Models: Theory, Evidence, and Scaling," *arXiv preprint arXiv:2511.02091*, 2025.

[4] S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski, "Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 20–22 Apr 2017, pp. 914–922.

[5] A. Doucet, N. de Freitas, K. Murphy, and S. Russell, "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'00. San Francisco, CA, USA: Morgan Kaufmann, 2000, pp. 176–183.

[6] L. Da Costa, P. Lanillos, N. Sajid, K. Friston, and S. Khan, "How Active Inference Could Help Revolutionise Robotics," *Entropy*, vol. 24, no. 3, p. 361, 2022. [Online]. Available: https://doi.org/10.3390/e24030361

[7] R. Tedrake and the Drake Development Team, "Drake: Model-Based Design and Verification for Robotics," https://drake.mit.edu, 2019, accessed: 2026-05-11.

[8] C. Heins, B. Millidge, D. Demekas, B. Klein, K. Friston, I. D. Couzin, and A. Tschantz, "pymdp: A Python library for active inference in discrete state spaces," *Journal of Open Source Software*, vol. 7, no. 73, p. 4098, 2022.

[9] S. Linderman, B. Antin, D. Zoltowski, and J. Glaser, "SSM: Bayesian Learning and Inference for State Space Models," https://github.com/lindermanlab/ssm, 2020.