Human-to-Robot Imitation with Symbolic Planning in a Unified Latent Space

Ruidong Ma
School of Computing
Sheffield Hallam University
Sheffield United Kingdom
ruidong.ma@shu.ac.uk

Wenjie Huang
Department of Computer Science
University of Manchester
Manchester United Kingdom
wenjie.huang@manchester.ac.uk

Zhegong Shangguan
Department of Computer Science
University of Manchester
Manchester United Kingdom
zhegong.shangguan@manchester.ac.uk

Angelo Cangelosi
Department of Computer Science
University of Manchester
Manchester United Kingdom
angelo.cangelosi@manchester.ac.uk

Alessandro Di Nuovo
School of Computing
Sheffield Hallam University
Sheffield United Kingdom
a.dinuovo@shu.ac.uk

Abstract

Direct imitation of humans by robots offers a promising direction for remote teleoperation and intuitive task instruction, where a human can perform a task naturally and the robot autonomously interprets and executes it using its own embodiment. Existing methods often rely on close alignment between human and robot scenes. This prevents robots from inferring the intent of the task or executing demonstrated behaviors when the initial states mismatch. Hence, it poses difficulties for non-expert users, who may need domain knowledge to adjust the setup.

To address this challenge, we propose a neuro-symbolic framework that unifies visual observations, robot proprioceptive states, and symbolic abstractions within a shared latent space. Human demonstrations are encoded into this representation as predicate states. A symbolic planner can thus generate high-level plans that account for the different robot initial states. A flow matching module then synthesizes continuous joint trajectories consistent with the symbolic plan.

We validate our approach on multi-object manipulation tasks. Preliminary results show that the framework can infer human intent and generate feasible symbolic plans and robot motions under mismatched initial states. These findings highlight the potential of neuro-symbolic models for more natural human-robot instruction. and they can enhance the explainability and trustworthiness of robot actions.

CCS Concepts

• **Computing methodologies** → **Learning from demonstrations**; **Robotic planning**.

Keywords

Human-to-Robot Imitation, Neuro-Symbolic Planning, Multimodal Learning

HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794399

ACM Reference Format:
Ruidong Ma, Wenjie Huang, Zhegong Shangguan, Angelo Cangelosi, and Alessandro Di Nuovo. 2026. Human-to-Robot Imitation with Symbolic Planning in a Unified Latent Space. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794399

1 Introduction

Human-to-robot imitation aims to enable robots to learn directly from human demonstrations despite differences in morphology and sensing. This capability is increasingly important for Human–Robot Interaction (HRI), where users expect to teach robots through natural demonstrations rather than specialized teleoperation interfaces. Recent work shows that robots can learn transferable skills from human videos [9] or aligned trajectories [4]. Once trained, models can allow the user to remotely teleoperate and instruct robot execution [11, 19].

However, if we consider a scenario where non-expert users instruct the robot to perform a manipulation task, which can contain various steps, they often demonstrate tasks in natural and unconstrained ways. Robots are typically required to be carefully reset or aligned with the demonstration, placing a significant burden on users. When the robot's initial state differs from the human demonstration, direct motion replay fails, and users must manually adjust the setup or provide additional guidance. This gap highlights the need for robot systems that can interpret demonstrations at a high-level with adaptive plans for new conditions, and behave in an interpretable way. Such capabilities directly support core HRI goals, including reducing user effort, increasing predictability, and improving trust during interaction.

In this work, our aim is to address this challenge by drawing inspiration from how humans understand and reproduce actions. Humans understand actions by inferring goals and object relations rather than copying movements [5]. Motivated by this goal-directed perspective, our approach represents demonstrated goals using symbolic abstractions. Given the robot data, we develop a multimodal internal model that can unify visual inputs, proprioception, and symbolic states into a shared latent space. Thus, a Planning

Domain Definition Language (PDDL) planner can generate a symbolic plan to connect the robot state to human demonstrations. Afterwards, a flow matching module can convert this plan into executable joint trajectories within the same latent space. This design enables context-aware imitation even when the initial state of the robot differs from the human demonstration during testing.

Our preliminary experiments in multi-object manipulation demonstrate that the robot shows generalization in resolving mismatches between the human demonstration and its own initial state.

Our contributions are summarized as follows:

(1) A unified framework that integrates visual observations, proprioception, and predicate-level abstractions into a shared internal model. This enables bidirectional grounding between symbolic task representations and robot embodiment.
(2) A neuro-symbolic planning pipeline that infers symbolic states from demonstrations. It generates high-level plans with PDDL and produces visual rollouts and joint trajectories through latent-space interpolation.
(3) A real-world multi-object manipulation study showing that our approach supports the robot in imitating from human videos under mismatched initial states. It improves interpretability via simulated visual trajectories.

2 Related work

Human-to-robot imitation seeks to transfer human behavior to robots despite the differences in morphologies. One direction is to focus on motion retargeting, where human bodies are mapped to robot kinematics through pose detection and inverse kinematics [4]. These approaches have enabled compelling whole-body and teleoperation-style behaviors. They can be further refined with reinforcement learning for improved tracking fidelity [1, 3, 21]. Recent extensions integrate vision–language models to incorporate object-aware cues into the retargeting process [11]. While highly effective for motion-focused imitation, these approaches often lack the capability to reason over long-horizon manipulation structure.

Recent efforts also aim to help robots understand human manipulation behaviours through more accessible demonstration modalities. Some works have focused on directly synthesizing robot action policies from tracked human hands[6, 10, 16]. Human demonstrators can teach the robot manipulation rules by editing visual scenes [7]. A robot can also actively ask humans guidance along with the increasing task complexity [15]. To further enable robots to imitate long-horizon tasks, recent studies can extract high-level and transferable skill abstractions from video. For instance, automatic skill discovery through clustering [19] and universal skill embeddings for robot action prediction via diffusion models [9] have been proposed. Similarly, human demonstration videos can be conditioned on language instructions to decompose trajectories into reusable skill units [20]. Although these approaches effectively enable intuitive teaching, they often focus on end-effector-level motion generation. Moreover, they rely on motion-level cues and often ignore the task-relevant structure that humans naturally infer, such as object relationships or causal dependencies. As a result, they lack interoperability of why the action is taken.

A complementary research direction draws inspiration from cognitive theories of human imitation. Models grounded in the mirror neuron system [17] can couple perception with action representations [14, 18, 22]. Thus, they can perform bidirectional mapping between observed behaviors and motor plans. The study in [2] shows that decomposing demonstrated skills into abstract cognitive units can support high-level reasoning and link motion primitives. Moreover, the neuro-symbolic approaches can derive symbolic abstractions from human demonstrations and therefore enable robots to adaptively approach goals by using action schemas. They show that high-level predicates and transition rules can be learned from only a few examples. It enables interpretable and efficient task generalization [8, 13]. However, these approaches are often focused on object information with extra detection modules, and often assume robot actions are known.

In summary, prior work on learning from human demonstrations has made imitation more accessible, yet these methods generally operate at the motion or skill level. They do not infer the underlying task structure. This limits interpretability and makes it difficult to adapt demonstrations to scenes that differ from original setup. Our framework addresses this limitation by inferring symbolic-level structure directly from raw human videos. This enables the robot to generate human-interpretable plans and to replan effectively under initial state mismatches.

On the other hand, current neuro-symbolic approaches provide interpretability through symbolic abstractions, but they typically require object-centric inputs. By grounding visual observations, proprioception, and symbolic predicates in a shared latent space, our system bridges the gap between raw video imitation and neuro-symbolic task reasoning. It allows symbolic plans to be decoded directly into robot-specific joint motions. This unified formulation enables adaptive imitation across embodiments.

3 Methodology

3.1 Symbolic Plan from human demonstration

We consider multi-object manipulation tasks, where there are multiple human videos representing different pick-and-place demonstrations. They are represented as symbolic states, which are annotated as sets of predicates and encoded as both object-object and human-object interactions. The predicate $on(x, y)$ represents a spatial relation between objects x and y, while $clear(x)$ indicates that the top surface of object x is free. $holding(x)$ means that the human hand is grasping an object without lifting it, and $free(g)$ states that the human hand is not holding any object. The corresponding robot data are defined in the same way in order to pair with human demonstrations. To reduce human labeling burden, we use GPT-4 to annotate them with human verification.

A compact PDDL thus specifies domain manipulation primitives with preconditions and effects. For example, in a complete pick-and-place sequence, the "grasp" primitive establishes the grasp by transforming the symbolic state from $\{free(g),\ clear(x)\ ,\ on(x, table)\}$ to $\{holding(x)\}$. The "place" primitive then relocates the object by requiring the predicates $\{clear(y)\ , on(y, table)\}$ and updates the state to $\{on(x, y),\ holding(x)\}$. The sequence is completed by a "release" primitive that removes $holding(x)$ and restores $free(g)$.

Although the robot is trained only on individual demonstrations, the PDDL-derived skills allow it to flexibly construct long-horizon task plans under different initial conditions.

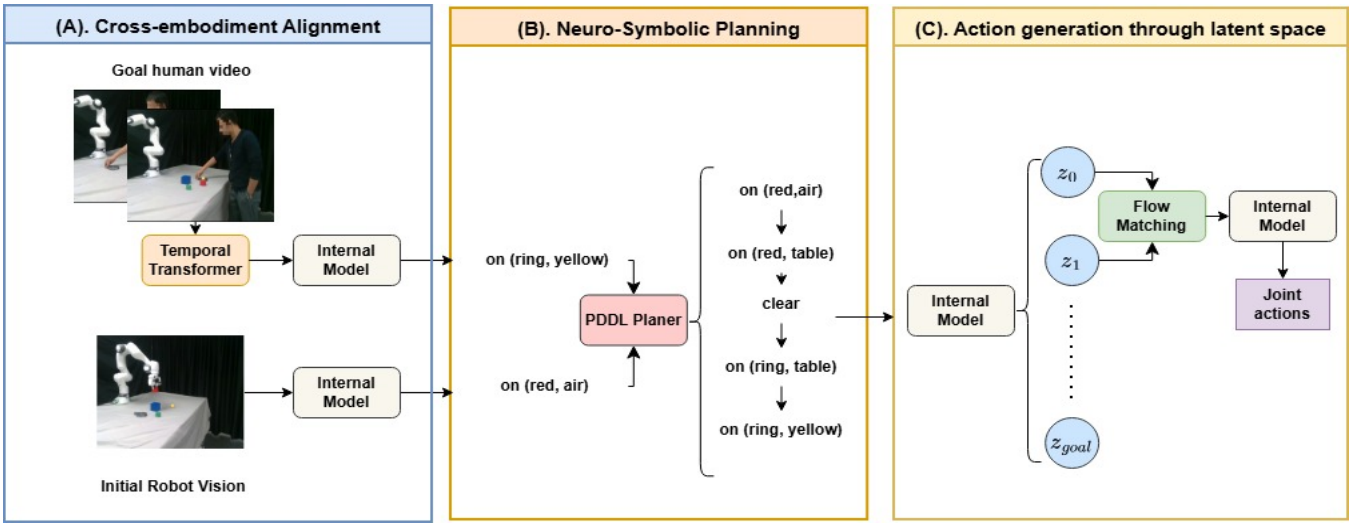


Figure 1: Overall Pipeline: (a) Human video demonstrations are mapped into the trained robot internal model. (b) The internal model infers predicate states and generates a symbolic plan using the PDDL planner. (c) These predicate states are then projected back into the latent space, where the flow matching module produces latent trajectories connecting successive states.

3.2 Robot Internal Model

We adopt a multimodal variational autoencoder (MVAE) as the robot's internal model I by integrating three modalities $O = \{v, j, s\}$, where v is the static robot image, j is the joint configuration and gripper states, and s is the corresponding one-hot encoded predicates.

Each modality is first encoded independently into latent variables: $z_v = E_v(v)$, $z_j = E_j(j)$, and $z_s = E_s(s)$. These latent codes are concatenated and processed by multi-layer-perceptrons (MLPs) to form a shared latent representation z. z can be further processed by MLPs and then divided back into modality-specific components as: $\hat{v} = D_v(z_v), \hat{j} = D_j(z_j), \hat{s} = D_s(z_s)$

To improve cross-modal robustness, we adopt a denoising training approach [22] by masking modalities with constant noise (e.g., $O = \{v, N, N\}$ or $O = \{N, s, N\}$). This encourages the model to reconstruct the full modalities from the partial inputs.

After training, the MVAE learns bidirectional cross-modal associations. Vision inputs can recover joint and predicate states, while predicates can reconstruct images and joints:

$$I(v) \mapsto \{\hat{j}, \hat{s}\}, \qquad I(s) \mapsto \{\hat{v}, \hat{j}\} \tag{1}$$

These mappings provide the foundation for both real-time imitation and internal simulation within our framework.

Moreover, we consider natural human demonstrations in which the demonstrator is not required to stand in the same physical location as the robot as shown in Fig.1. To bridge this discrepancy, we align human demonstrations and robot visions at the predicate level. The human videos are encoded by a temporal transformer T. It is trained by minimising the difference between the human embedding z_h and the corresponding robot embedding z_r from the internal model:

$$z_h = T(h), \qquad z_h \approx z_r, \tag{2}$$

3.3 Action generation through latent space

The internal model can therefore process the goal human video and the initial robot image into predicate-level states. Based on these states, the PDDL planner can produce a sequence of symbolic states, which are projected back into the internal model's latent space as: $\{z_0, z_1, \ldots, z_g\}$.

In this sequence, each latent state z_{k+1} acts as the local goal for the current state z_k. To connect these states, we generate a latent trajectory using a flow matching module F [12]. Each transition is modeled as a goal-conditioned interpolation:

$$\hat{\mathcal{Z}}_k = F(z_k | z_{k+1}) \tag{3}$$

The flow matching module is trained by supervising the velocity that drives latent trajectory toward their goals. Each latent trajectory is donated as $\hat{\mathcal{Z}}_k = \{\mathcal{Z}_k[1], \mathcal{Z}_k[2], \ldots, \mathcal{Z}_k[t]\}$ with time step t [12].

Finally, each generated latent element is decoded into robot image and joints configuration as:

$$(\hat{v}[t],\ \hat{j}[t]) = I(\hat{\mathcal{Z}}[t]). \tag{4}$$

The internal simulated visual trajectories can be used to interpret the planning process, while the joints are used for joint position control in real world.

4 Experiment and result

4.1 Experiment Setup

We evaluate the framework on a real-world multi-object tabletop manipulation task involving three cubes, a ball, and a ring. In this setup, the human can place either the red cube, green cube, or yellow ball onto the blue cube, and place the hexagonal ring onto either the green cube or the yellow ball. The Franka fr3 arm is used.

In this preliminary experiment, we have two human demonstrators with limited knowledge of robotics. For each task, we record

Table 1: Average SR versus the number of sub-tasks.

	Num. of Sub-tasks				
	One	Two	Three	Four	Five
Ours	0.98	0.98	0.97	0.96	0.96
LFM	0.97	0.97	0.08	0.00	0.00

one human and one robot demonstration. In contrast to prior works [9, 19], the human demonstrators operate from a different physical location, providing a more natural and unconstrained setting.

Keyframes are annotated using a predefined predicate sets. Additionally, intermediate frames in which the hand is holding an object are labeled as *on(obj, air)*, and approaching motions are annotated with the extra predicate *approach*. These predicates define the PDDL planning domain.

The internal model and flow matching module are trained on individual manipulation demonstrations. At test time, the robot begins with states that differ from human demonstration. The evaluation examines whether the robot can infer meaningful predicates and thus generate long-horizon plans with joint control signals.

4.2 Result

In preliminary experiments, we evaluate our framework's ability to imitate human demonstrations with different numbers of subtasks. One subtask corresponds to a pick or place action, two form a full pick–place sequence, and three or more require generalization to resolve mismatches (Fig. 2). We compare against an ablation baseline Latent Flow Matching (LFM), where the robot is directly controlled by flow matched latent without planning. It is trained with separate pick-place demonstrations similar to our framework.

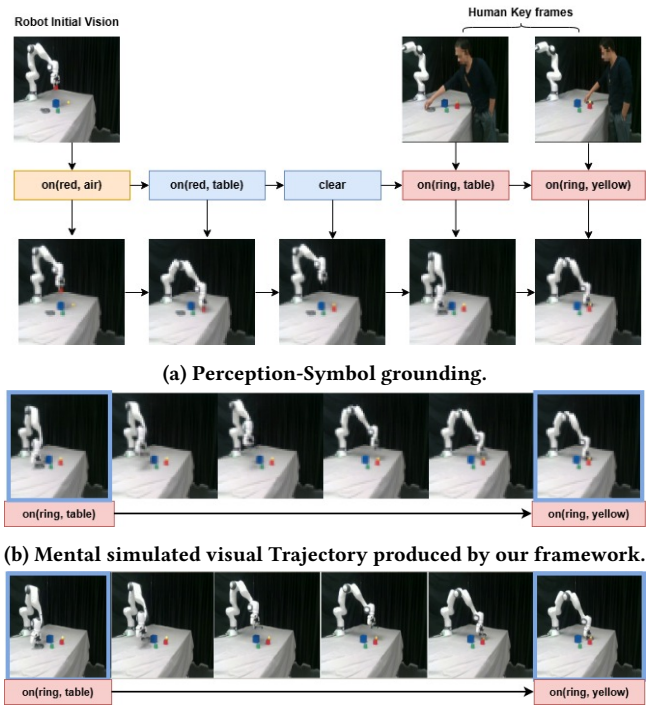
In each case, we use 20 randomly sampled robot initial states against randomly selected human demonstrations. The success rates (SRs) over five trials are reported to assess whether the robot completes the task consistent with the human demonstration. As described in Table.1, symbolic planning can improve the robot generalization under human and robot mismatches.

Fig.2 describes one representative task. The robot observes a demonstration that the ring is placed on the yellow ball. However, the robot is already holding a red cube, which makes it impossible to pick up the ring directly. The robot must therefore generate a new plan, such as releasing the cube first, as shown in Fig. 2a. The intermediate latent trajectories can be decoded back into visual predictions shown in Fig. 2b, as well as coherent full-joint trajectories used to control the real robot shown Fig. 2c. More experiments can be found in the companion video.

As our framework integrates the modalities in a unified latent space, the robot's behaviour becomes easy to interpret by visualizing the latent trajectories as shown in Fig. 2a and 2b. This transparency has potential for more human-friendly imitation by allowing users to understand and trust how the robot will execute a demonstrated task.

5 Conclusion

We introduced a neuro-symbolic framework that enables robots to interpret human demonstrations, generate symbolic plans, and produce continuous joint motions within a unified latent space. Our



(a) Perception-Symbol grounding.

(b) Mental simulated visual Trajectory produced by our framework.

(c) Real-world joints control produced by our framework.

Figure 2: (a) The internal model grounds both robot and human visual inputs into predicate states (red boxes), and the PDDL planner generates the corresponding symbolic sequence (blue boxes). These symbolic states are projected back through the internal model to predict visual observations (third row). (b) Thus, the imagined visual trajectories can be produced by the flow matching module. (c) And the corresponding joint trajectories are executed by the robot. We only show the representative predicates.

primarily results show that the robot can imitate human demonstrations even when its initial state does not match the scene, and can visualize intermediate plans in a way that is interpretable to users. These capabilities reduce user burden. It also shows the potential to improve the transparency of robot behaviour and thus strengthen usability and trust, especially for no-expert users.

Future work will extend the framework to more complex manipulation tasks and include broader comparative evaluations. We plan to conduct user studies to evaluate how well the system supports non-expert teaching, perceived transparency, and trust. In addition, the current system relies on a manually defined PDDL domain. To further reduce human effort, we plan to explore solutions to automatically construct such symbolic domains.

Acknowledgments

This work is funded by the Innovate UK (grant number 10089807 for the Horizon Europe project PRIMI Grant agreement n. 101120727).

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. 2022. Human-to-Robot Imitation in the Wild. In *Robotics: Science and Systems*.

[2] Zixuan Chen, Ze Ji, Shuyang Liu, Jing Huo, Yiyu Chen, and Yang Gao. 2024. CasIL: Cognizing and Imitating Skills via a Dual Cognition-Action Architecture. In *23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[3] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. 2024. Expressive Whole-Body Control for Humanoid Robots. In *Robotics: Science and Systems*.

[4] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. 2024. HumanPlus: Humanoid Shadowing and Imitation from Humans. In *Conference on Robot Learning (CoRL)*.

[5] Marc Jeannerod. 2001. Neural Simulation of Action: A Unifying Mechanism for Motor Cognition. In *NeuroImage*, Vol. 14. Academic Press Inc. doi:10.1006/nimg.2001.0832

[6] Ananth Jonnavittula, Sagar Parekh, and Dylan P. Losey. 2024. VIEW: Visual Imitation Learning with Waypoints. (April 2024).

[7] Karthik Mahadevan, Blaine Lewis, Jiannan Li, Bilge Mutlu, Anthony Tang, and Tovi Grossman. 2025. ImageInThat: Manipulating Images to Convey User Instructions to Robots. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2025*. IEEE.

[8] Leon Keller, Daniel Tanneberg, and Jan Peters. 2025. Neuro-Symbolic Imitation Learning: Discovering Symbolic Abstractions for Skill Learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 6519–6526. arXiv:2503.21406 [cs] doi:10.1109/ICRA55743.2025.11127692

[9] Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. 2025. UniSkill: Imitating Human Videos via Cross-Embodiment Skill Representations. In *9th Conference on Robot Learning (CoRL 2025), Seoul, Korea*.

[10] Marion Lepert, Jiaying Fang, and Jeannette Bohg. 2025. Phantom: Training Robots Without Robots Using Only Human Videos. In *9th Conference on Robot Learning*.

[11] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. 2024. OKAMI: Teaching Humanoid Robots Manipulation Skills through Single Video Imitation. In *8th Conference on Robot Learning (CoRL 2024)*.

[12] Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. FLOW MATCHING FOR GENERATIVE MODELING. In *2023 International Conference on Learning Representations (ICLR)*.

[13] Pierrick Lorang, Hong Lu, Johannes Huemer, Patrik Zips, and Matthias Scheutz. 2025. Few-Shot Neuro-Symbolic Imitation Learning for Long-Horizon Planning and Acting. arXiv:2508.21501 [cs] doi:10.48550/arXiv.2508.21501

[14] Kristína Malinovská and Jakub Mišovský. 2024. Robotic Model of the Mirror Neuron System: A Revival. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 15025 LNCS. Springer Science and Business Media Deutschland GmbH, 313–323. doi:10.1007/978-3-031-72359-9_23

[15] Muhan Hou, Koen Hindriks, A.E. Eiben, and Kim Baraka. 2025. Active Robot Curriculum Learning from Online Human Demonstrations. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE.

[16] Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. 2024. R+X: Retrieval and Execution from Everyday Human Videos. (July 2024).

[17] Giacomo Rizzolatti and Leonardo Fogassi. 2007. Mirror Neurons and Social Cognition. In *Handbook of Evolutionary Psychology*. Oxford University Press.

[18] M. Yunus Seker, Alper Ahmetoglu, Yukie Nagai, Minoru Asada, Erhan Oztop, and Emre Ugur. 2022. Imitation and Mirror Systems in Robots through Deep Modality Blending Networks. *Neural Networks* 146 (Feb. 2022), 22–35. doi:10.1016/j.neunet.2021.11.004

[19] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. 2023. XSkill: Cross Embodiment Skill Discovery. In *7th Annual Conference on Robot Learning*.

[20] Ke Ye, Jiaming Zhou, Yuanfeng Qiu, Jiayi Liu, Shihui Zhou, Kun-Yu Lin, and Junwei Liang. 2025. From Watch to Imagine: Steering Long-horizon Manipulation via Human Demonstration and Future Envisionment.

[21] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. 2021. XIRL: Cross-embodiment Inverse Reinforcement Learning. In *5th Conference on Robot Learning (CoRL 2021), London, UK*.

[22] Martina Zambelli, Antoine Cully, and Yiannis Demiris. 2020. Multimodal Representation Models for Prediction and Control from Partial Information. *Robotics and Autonomous Systems* 123 (2020), 103312–103312. doi:10.1016/j.robot.2019.103312

Received 2025-12-08; accepted 2026-01-12