Towards Robust Estimation of Human Intention Hierarchy in Robot Teleoperation

Nikki Lijing Kuang*
University of California San Diego
l1kuang@ucsd.edu

Songpo Li
Honda Research Institute
songpo_li@honda-ri.com

Soshi Iba
Honda Research Institute
siba@@honda-ri.com

Abstract

The past few decades have witnessed the widespread adoption of robot teleoperation across various real-world domains such as manufacturing and healthcare. It has been recognized as an effective approach to assist humans in remotely tackling tasks that pose significant challenges and risks when undertaken alone. To improve the efficiency of collaboration between human and robot in teleoperated systems, it is crucial for the robot to precisely infer human intentions. In this work, we introduce RoHIE, a novel architecture designed to reason about the intentions of the human partner at different levels of granularity. RoHIE leverages non-verbal observations that capture the motion and gaze information in shared autonomy, and learns a flexible intention hierarchy to categorize the relationship between low-level action primitives and higher-level task goals, thereby enabling robust inference. Moreover, by learning a compact representation in the embedding space, our framework captures the latent structural information of human behaviors from human partners' demonstrations, empowering the robot to robustly and accurately estimate the intention of new human companions. Finally, we rigorously validate the efficacy of our framework on a teleoperation dataset consisting of a variety of building block assembly tasks.

1 Introduction

The rapid advancement and increasing sophistication of modern engineering present great challenges in fully automating robot manipulation or relying solely on human operators to manually perform tasks, especially in hazardous environments. In response, assistive robot teleoperation (Figure 1) has emerged as a compelling solution, facilitating successful human-robot interaction (HRI) by striking a balance between full autonomy and pure teleoperation across various domains. It assists humans in a spectrum of tasks, spanning from warehouse management, medical surgeries, to underwater operations and space missions. Arguably, it amalgamates the strengths of both worlds: leveraging the efficiency of robotic systems for repetitive structured tasks, while harnessing the intelligence and dexterity of human operators to address real-world problems.

Despite the promising paradigm of robot teleoperation, its deployment is significantly impacted by communication delays and the inherent disparities between local teleoperated sites and remote environments (Zhu et al., 2023). One of the fundamental challenges in human-robot teleoperation is the rapid and reliable inference of human intention (Wang et al., 2021; Li et al., 2020). Notably, accurate estimates of the intention of human partners allow remote robots to act proactively, thereby enhancing the robustness of robotic response against latency. However, existing methods of intention estimation often rely solely on the observations of human hands in the physical interaction between robots and human partners, which can be problematic when human hand position coincides with that of the robot endpoint (Lai et al., 2022).

*This work was done while the author was an intern at Honda Research Institute, USA.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

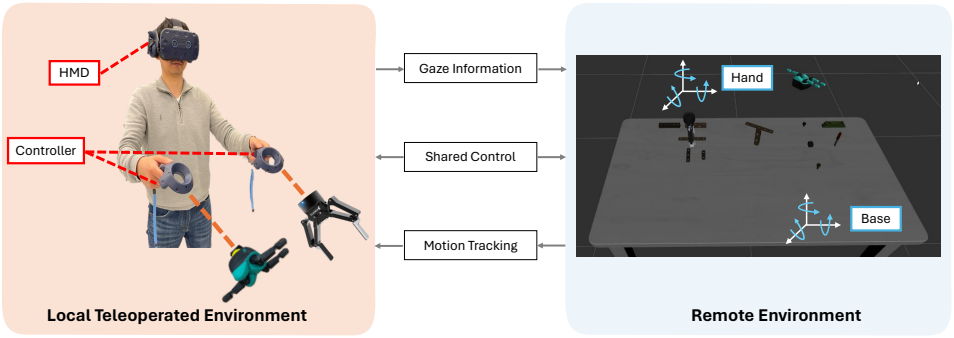

Figure 1: A shared-control human-robot teleoperation system with HMD and controller.

Moreover, the intention hierarchy presented in human-level intelligence often fails to be captured when estimating the goals of human collaborators, further impeding the estimation accuracy in teleoperated systems (Jain and Argall, 2019; Huang et al., 2023; Shen et al., 2023). In particular, the successful execution of complex tasks in a shared-control architecture hinges on meticulous step-by-step procedures, encompassing both higher-level task objectives and lower-level execution of sub-goals. Therefore, it is imperative to determine human intention at each hierarchical level. However, inferring higher-level intention is challenging, as immediate observations of motion and gaze data only afford timely insights into the lowest-level operations. Additionally, lower-level operations and intermediate sub-goals may share commonalities across different tasks, introducing ambiguity that further compromises the quality of inference pertaining to higher-level intentions.

To jointly address the above challenges in assistive robot teleoperation, we introduce RoHIE, a novel architecture to reason about the intentions of the human partner at different levels of granularity. RoHIE takes multiple forms of observations as input to enable contact-free intention estimation tailored for teleoperation. Specifically, we utilize behavioral data that can be easily captured and transmitted through non-verbal communication, including motion data that tracks the movement of robot hands and objects from the remote environment, as well as gaze data that captures the secondary human behaviors from the teleoperated environment without direct oral commands. Compared to model-based methods that require defining hierarchical structure apriori (Huang et al., 2023), RoHIE excels at automatically discovering the hidden structure of intention hierarchy from behavioral data, yielding exceptional flexibility in large-scale deployment. Such capability of auto-knowledge discovery allows our paradigm to achieve higher accuracy in downstream tasks.

Our main contributions are summarized as follows:

- We propose a novel architecture, RoHIE, to reason about the intentions of the human partner in assistive robot teleoperation through non-verbal observations.
- To ensure the interpretability and the robust inference of low-level action primitives, and to generalizate to unseen human behaviors, representation learning is used in conjunction with an action transition model to capture and refine the mutual behavior information in observation sequences while retaining the original semantics.
- We learn the human intention hierarchy through unsupervised hierarchical clustering to eliminate the need of a pre-defined model design, allowing flexible structure discovery in large-scale tasks with complex architecture.
- Empirical evaluations are performed on a teleoperation dataset to showcase the effectiveness of RoHIE, which covers a range of building block assembly tasks.

2 Problem Formulation

Consider the human-robot teleoperation setup where the human partner remotely controls robot arms to complete a task $h^* \in \mathcal{H}$ by manipulating a set of objects O. To complete the task h^*, at each time step t, the human needs to operate the robot to perform some low-level action $a_t^* \in \mathcal{A}$. Both the task space $\mathcal{H}$ and action space $\mathcal{A}$ are discrete and stationary. To allow seamless communication between human operator and the robot, we are interested in studying the intention inference problem

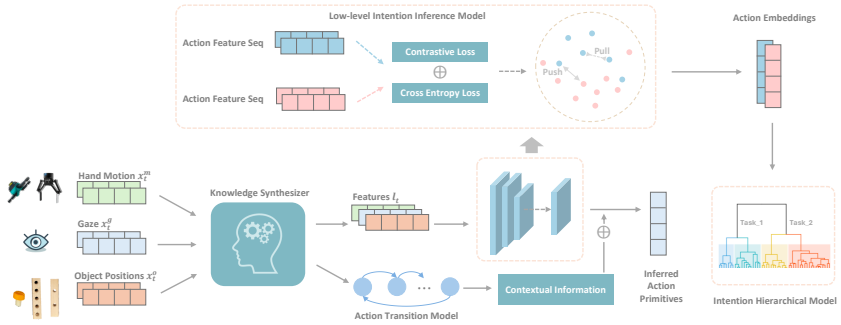

Figure 2: Overview of RoHIE architecture.

where the human's high-level task goal h^* and the low-level intended action a_t^* at each time step are unknown to the robot. Here, only specific forms of non-verbal data $X_t := [x_t^g, x_t^m, x_t^o] \in \mathbb{R}^{1\times d_r}$ can be observed by the robot, where $x_t^g \in \mathbb{R}^{1\times d_g}$ denotes the gaze of the human, $x_t^m \in \mathbb{R}^{1\times d_m}$ is the motion of the arms, and $x_t^o \in \mathbb{R}^{1\times d_o}$ incorporates the positional information of all objects $o \in O$. The robot needs to infer (predict) the intention hierarchy to complete the overall task as well as the most likely action a_t that the human partner is trying to perform. In general, the intention hierarchy I encapsulates the human intention at different levels of granularity (e.g. low-level actions, sub-procedures, high-level task) and the relationship between each intention level.

3 Robust Hierarchical Intention Estimation

We first present the model architecture of RoHIE (Figure 2). It is a novel paradigm that aims to address the intention inference problem with non-verbal observations in human-robot teleoperation, where the human operator remotely controls robot arms to complete a set of predefined tasks. It contains (1) a knowledge synthesizer for learning the behavioral features and action transition model from observations; (2) a low-level intention inference model for inferring the action primitives and their semantic latent representations; and (3) a hierarchical intention model for learning the human intention hierarchy. Algorithm 1 describes the generic algorithm of human intention estimation. Details of each major components are discussed in the remainder of this section.

Knowledge Synthesizer. It takes the motion, gaze, and positional data as input, and leverages domain expertise to transform and synthesize essential behavioral archetypes and geometric characteristics from these observations. By learning a mapping $\varphi : \mathcal{X}_r \to \mathcal{X}_l$ to extract feature $\widetilde{X}_t$ from temporal sequential observation at time t, it encapsulates the object being gazed at, the relative spatial position of each operatable object, and the orientation of the manipulated object, etc. Additionally, it learns a time-homogeneious action transition model $\pi : \mathcal{A} \to \Delta_{\mathcal{A}}$, where $\pi(a'|a)$ signifies the probability of taking action a' following action a. Given a sequence of actions $\{a_\tau\}_{\tau=1}^t$, the likelihood of this sequence occurring is computed as: $p(a_1, \ldots, a_t) = p(a_1)\prod_{\tau=1}^{t-1}\pi(a_{\tau+1}|a_\tau)$. If the inferred action sequence yields zero likelihood deduced by π, a new action primitive a_t is regenerated from the last layer of the model until it results in a valid sequence.

Low-level Intention Inference Model. We define an *intention-aware loss* $\mathcal{L}_{tl}$ as a mixture of cross-entropy (CE) loss $\mathcal{L}_{ce}$ and contrastive loss $\mathcal{L}_{wnce}$:

$$\mathcal{L}_{tl} = (1-\lambda)\mathcal{L}_{ce} + \lambda\mathcal{L}_{wnce}, \tag{1}$$

where λ is an adaptive hyperparameter being adjusted iteratively to ensure the best performance. Specifically, an RNN-based model is trained using synthesized features $\{\widetilde{X}_t\}_{t=1}^{|D_{train}|}$ with a cross-entropy loss to maximize the conditional log-likelihood: $\mathcal{L}_{ce} = -\sum_{t=1}^{|D_{train}|}\log p_\theta(a_t^*|\widetilde{X}_t)$. To capture the underlying patterns human behaviors, we minimize a weighted supervised InfoNCE loss $\mathcal{L}_{wnce}$ for the latent embeddings of the customized RNN:

$$\mathcal{L}_{wnce} = -\sum_{i=1}^{|D_{train}|}\log\frac{\exp(s_{i,i})}{\sum_{a_i^*\neq a_j^*}\exp(s_{i,j}) + \alpha\sum_{a_i^*=a_j^*}\exp(s_{i,j})}, \tag{2}$$

where α is a weighted hyperparameter. Intuitively, optimizing our model with $\mathcal{L}_{tl}$ encourages learning latent patterns of human behavior in the embedding space, preserving the temporal coherence, sematic interpretability, and mutual information of low-level intention.

Algorithm 1: Robust Intention Estimation in Human-Robot Teleoperation

Input: Collected teleoperated data $D_{train} = \{X_t, a_t^*, h^*\}_{t=1}^{T_1}$, $D_{test} = \{X_t\}_{t=1}^{T_2}$

Output: low-level action intention $\{a_t\}_{t=1}^{T_2}$, intention hierarchies $\{I_h\}_{h\in\mathcal{H}}$ for each task $h \in \mathcal{H}$

Knowledge synthesis:
$\varphi(X_t) \rightarrow \widetilde{X}_t,\ \forall X_t \in D_{train}, D_{test}$;
Learn transition model π (Equation (3)) on D_{train};
for *epoch* $i = 1, \ldots, K$ **do**
  for *batch* $b = 1, \ldots, B$ **do**
    Train low-level action inference model using intention-aware loss $\mathcal{L}_{tl}$ (Equation (7));
Infer action intentions $\{a_t\}_{t=1}^{T_2}$ and their latent representation $\{\widetilde{Z}_t\}_{t=1}^{T_2}$ for D_{test};
Refine action inference $\{a_t\}_{t=1}^{T_2}$ with π;
Learn intention hierarchy $\{I_h\}_{h\in\mathcal{H}}$ for each task;
return $\{a_t\}_{t=1}^{T_2}, \{I_h\}_{h\in\mathcal{H}}$;

Structure Discovery of Intention Hierarchy. We employ unsupervised hierarchical clustering to automatically discover the intermediate sub-procedures in the intention hierarchy. At each level of the intention hierarchy, sequences corresponding to the same sub-procedure are grouped into one cluster using Agglomerative clustering, based on the similarity measured by the Ward's linkage. RoHie thus forms an interpretable hierarchical intention structure in an unsupervised fashion.

4 Experiments

In this section, we assess the performance of our framework in terms of the quality of inference regarding low-level action intention, the capability to generalize and infer intentions of new human partners, the model quality with respect to data efficiency, and the interpretability of the resulting intention hierarchy. More details can be found in Appendix C.

Environment Setpup. We design a teleoperated experimental environment using virtual reality (VR) technology. The human operator interfaces with the environment through a HTC Vive Pro Eye headset, which provides immersive visualization of the remote operating environment rendered by the rviz 3D ROS framework. To interact with the virtual remote environment, the operator employs two Oculus controllers, each dedicated to controlling a distinct virtual robot arm. Furthermore, we employ a binocular Tobii eye tracker operating at 120 Hz to track the human operator's gaze. Object positions are tracked through stereo cameras.

Block-assembly Dataset. We invited thirteen human participants to engage in a series of six meticulously-designed building block assembly tasks (Figure 3) within the aforementioned teleoperated environment. Details of each building block assembly task can be found in Table 2. Notably, certain objects may be utilized across multiple tasks, leading to the recurrence of identical action primitives in different assembly tasks. Across these tasks, a total of 194 valid trials are collected, with each representing a unique instance of task execution by the participants. The average duration of each trial is 90 time seconds.

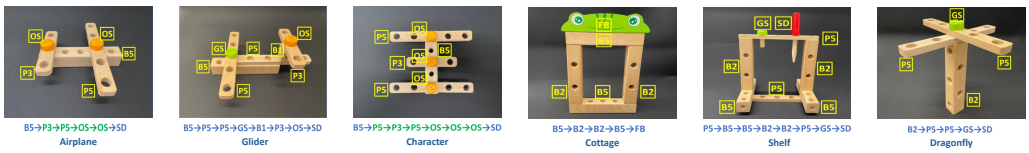

Figure 3: Visualization of building block assembly tasks, with objects labeled in yellow. Each task is accompanied by a workflow outlining a correct sequence of operations required to complete the task. Objects highlighted in green can be rearranged in different orders.

Inference of Low-level Action Primitives. As shown in Table 1 (left), simple baselines such as randomly inferring action primitives or selecting the most frequent action yield notably poor performance. While classical machine learning baselines and deep neural network baselines show similar performance, RoHIE consistently outperforms them across all evaluation metrics.

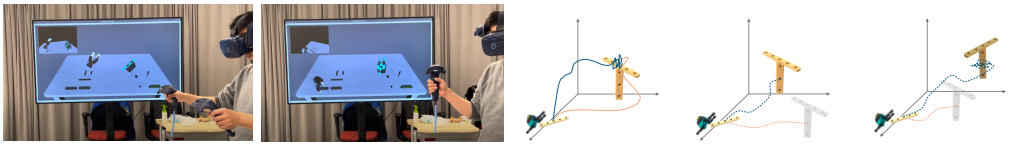

Figure 4: Illustration of different human behaviors in "dragonfly" building-block assembly task. Left: actual human teleoperation. Right: different human hand motion trajectories: (1) moving objects along different spatial regions; (2) placing assembled blocks to different relative locations; (3) rotating building blocks to adjust operating views while keeping the same geometric structure.

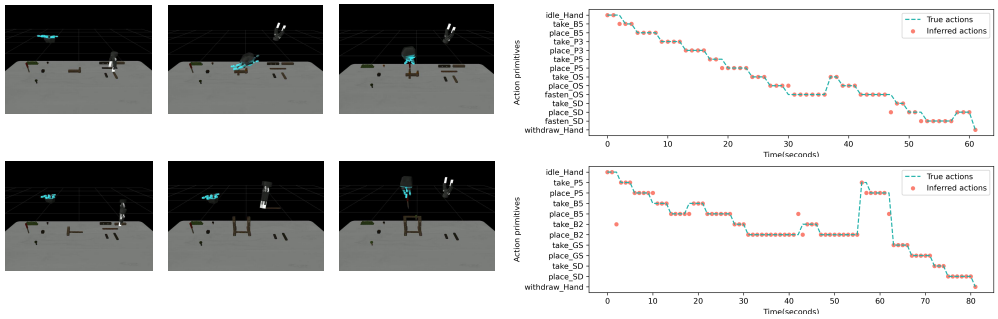

Figure 5: RoHIE action intention inference of a new human partner in completing different block assembly tasks. Top: intention inference when performing the"airplane" assembly task. Bottom: intention inference when performing the "shelf" assembly task.

	Acc.	**Prec.**	**Recall**	**F1**
Random	4.13	5.57	4.13	4.46
Majority	10.48	1.10	10.48	1.99
NaiveBayes	75.95	78.85	75.95	75.08
SVM	78.37	81.03	78.37	78.17
DecisionTree	75.83	77.97	75.83	76.02
RandomForest	78.69	79.12	78.69	77.08
MLP	77.51	80.49	77.51	76.80
RNN_baseline	77.36	77.53	77.36	77.20
RoHIE	**90.96**	**87.39**	**90.96**	**86.98**

N_{tu}	11	9	7	5
$\|T_{train}\|$	167	124	97	68
$\|T_{test}\|$	27	70	97	126
RandomForest	80.32	78.69	73.79	67.69
RNN_baseline	77.69	77.36	70.51	70.26
RoHIE	**90.29**	**90.96**	**91.35**	**88.78**

Table 1: Left: Inference result of low-level action primitives on 70 testing trails that are performed by 4 new human users who have not participated in the training. Right: Inference result with varied sizes of training trails, where N_{tu} specifies the number of human operators involved during training, $|T_{train}|$ and $|T_{test}|$ denote the number of trails for training and testing respectively. Results are reported in percentage (%). Abbreviations: Acc. = Accuracy, Prec. = Precision.

Generalization and Robustness. Human participants may demonstrate substantially different behaviors (Figure 4) when teleoperating robot arms. As shown in Figure 5, RoHIE demonstrates effective generalization capability in accurately inferring the intentions of new human partners by consistently eliminating irrational action primitives unrelated to the ongoing assembly task. Additionally, we evaluate the robustness of its performance in Table 1 (right). It demonstrates that RoHIE maintains high inference accuracy even with a small number of training trials, whereas the performance of the Random Forest algorithm and RNN baseline rapidly deteriorates as the size of the training dataset decreases.

5 Discussions and Conclusion

We introduce a novel paradigm RoHIE to reason about the human intentions at different granularity in assistive robot teleoperation. It not only provides robust estimation of low-level action primitives at each time step, but also captures the useful semantic latent representation from temporal observations of gaze and motion data without requiring verbal interaction with human. Building upon the low-dimensional action embeddings, it learns a flexible intention hierarchy through unsupervised hierarchical clustering, eliminating the need of domain-specific knowledge to define the structure of intention hierarchy apriori.

References

Amor, H. B., Neumann, G., Kamthe, S., Kroemer, O., and Peters, J. (2014). Interaction primitives for human-robot cooperation tasks. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 2831–2837. IEEE.

Campbell, J., Stepputtis, S., and Amor, H. B. (2019). Probabilistic multimodal modeling for human-robot interaction tasks. *arXiv preprint arXiv:1908.04955*.

Carmichael, M. G., Aldini, S., Khonasty, R., Tran, A., Reeks, C., Liu, D., Waldron, K. J., and Dissanayake, G. (2019). The anbot: An intelligent robotic co-worker for industrial abrasive blasting. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8026–8033. IEEE.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Darvish, K., Penco, L., Ramos, J., Cisneros, R., Pratt, J., Yoshida, E., Ivaldi, S., and Pucci, D. (2023). Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics*.

Gao, T., Yao, X., and Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Hamaya, M., Matsubara, T., Noda, T., Teramae, T., and Morimoto, J. (2017). Learning assistive strategies for exoskeleton robots from user-robot physical interaction. *Pattern Recognition Letters*, 99:67–76.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Huang, Z., Mun, Y.-J., Li, X., Xie, Y., Zhong, N., Liang, W., Geng, J., Chen, T., and Driggs-Campbell, K. (2023). Hierarchical intention tracking for robust human-robot collaboration in industrial assembly tasks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9821–9828. IEEE.

Jain, S. and Argall, B. (2019). Probabilistic human intent recognition for shared autonomy in assistive robotics. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(1):1–23.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Koert, D., Pajarinen, J., Schotschneider, A., Trick, S., Rothkopf, C., and Peters, J. (2019). Learning intention aware online adaptation of movement primitives. *IEEE Robotics and Automation Letters*, 4(4):3719–3726.

Lai, Y., Paul, G., Cui, Y., and Matsubara, T. (2022). User intent estimation during robot learning using physical human robot interaction primitives. *Autonomous Robots*, 46(2):421–436.

Li, S., Bowman, M., Nobarani, H., and Zhang, X. (2020). Inference of manipulation intent in teleoperation for robotic assistance. *Journal of Intelligent & Robotic Systems*, 99:29–43.

Losey, D. P., McDonald, C. G., Battaglia, E., and O'Malley, M. K. (2018). A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. *Applied Mechanics Reviews*, 70(1):010804.

Shen, Y., Mo, X., Krisciunas, V., Hanson, D., and Shi, B. E. (2023). Intention estimation via gaze for robot guidance in hierarchical tasks. In *Annual Conference on Neural Information Processing Systems*, pages 140–164. PMLR.

Singh, S., Arora, C., and Jawahar, C. (2016). First person action recognition using deep learned descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2620–2628.

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. (2021). Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*, pages 9870–9879. PMLR.

Wang, W., Li, R., Chen, Y., Sun, Y., and Jia, Y. (2021). Predicting human intentions in human–robot hand-over tasks through multimodal learning. *IEEE Transactions on Automation Science and Engineering*, 19(3):2339–2353.

Yan, A., He, Z., Li, J., Zhang, T., and McAuley, J. (2023). Personalized showcases: Generating multi-modal explanations for recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2255.

Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., and Hsu, C.-N. (2021). Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*.

Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., and Huang, J. (2022). Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Zhu, Y., Fusano, K., Aoyama, T., and Hasegawa, Y. (2023). Intention-reflected predictive display for operability improvement of time-delayed teleoperation system. *ROBOMECH Journal*, 10(1):1–11.

Appendices

A Related Work

Robot teleoperation. Recent advances in collaborative robots have boosted the prevalence of HRI applications in both industrial and domestic environments, which emphasizes the teamwork and shared control between robots and humans in solving domain-specific problems collaboratively Zhu et al. (2023). Robotic teleoperation is esteemed for its capability to remotely perform tasks that involve complex human decision-making and procedures that can be executed by human operators at a distance Darvish et al. (2023). It offers exclusive advantages over fully automated robots by combining robotic autonomy and human supervision with intelligent reasoning. It allows both parties to focus on the aspects that they are best adept at, in a blend that can entail different degrees of autonomy for the robotic part Losey et al. (2018). Specifically, different subtasks are assigned such that the remote robot is responsible for detailed low-level operations, such as lower kinematic aspects of action execution, while the human teleoperator focuses on high-level decision-making, such as task planning and exception handling. To improve the robustness of human-robot interaction, the ability to learn, generalize, and adapt to different tasks is crucial. While task-oriented measures are commonly used to address the human behavior variability in HRI applications Carmichael et al. (2019); Hamaya et al. (2017), such measures may degrade the interactions for non-expert users who do not have prior knowledge of the system dynamics. In comparison, our framework is capable of generalizing to new human behaviors using representation learning to capture the structural commonalities from different users.

Intention estimation. Prior works of intention estimation in HRI have primarily focused on interaction primitives Amor et al. (2014), which utilize observations of human hands to infer intention and generate suitable robotic trajectories in response. Along this line of work, extensions are developed by utilizing human-robot gestures Amor et al. (2014), hand shaking Campbell et al. (2019), interaction forces Lai et al. (2022), and movement primitives Koert et al. (2019). However, such methods are infeasible in contact-free teleoperated environments. In such settings, relying solely on hand positions for end-to-end action recognition proves inadequate, as hand position and robot endpoint locations may overlap during movement. Additionally, egocentric vision approaches Singh et al. (2016) require calibration of human operators' viewpoints and fail to generalize to different scenes. In contrast, our work jointly utilizes motion data, gaze information captured by stereo cameras as well as object positional data.

Representation learning. Contrastive learning has emerged as a powerful technique for representation learning, with applications ranging from computer vision He et al. (2020); Yan et al. (2021), language models Yang et al. (2022); Yan et al. (2023) to reinforcement learning (RL) Stooke et al. (2021). It functions by reducing the distance between similar data points within an embedding space while simultaneously increasing the separation between dissimilar data points through the usage of contrastive losses, which are at the core of the recent advances in representation learning. These losses are usually comprised of an attractive term and a repulsive one, where the former guarantees similar samples to have close representations while the latter pushes representations of dissimilar samples far apart. Recent work by Khosla et al. Khosla et al. (2020) introduced contrastive multiview coding, a variant of contrastive learning, for learning latent representations of human actions. By considering multiple views of the same action data, this approach enhances the model's ability to capture fine-grained details and temporal dependencies. Besides, frameworks such as SimCLR Chen et al. (2020), MoCo Yang et al. (2022), and SimCSE Gao et al. (2021) exhibit promising performance in zero-shot and few-shot learning. Likewise, our framework relies on the recent advancements in representation learning.

B Details of RoHIE Architecture

B.1 Knowledge Synthesizer

In teleoperated environments, the precise inference of human intentions relies heavily on identifying and interpreting human behavioral information from available observations. The knowledge synthesizer within RoHIE serves as a pivotal tool for distilling meaningful insights from various forms of non-verbal observations. Specifically, it takes the motion, gaze, and positional data as in-

put, and leverages domain expertise to transform and synthesize essential behavioral archetypes and geometric characteristics from these observations.

It first learns a mapping $\varphi : \mathcal{X}_r \to \mathcal{X}_l$ to extract multidimensional feature $\widetilde{X}_t$ from temporal sequential observation at time t, where $\mathcal{X}_r \in \mathbb{R}^{d_r}$ represents the space of input observation and $\mathcal{X}_l \in \mathbb{R}^{d_l}$ denotes the high-dimensional feature space. Notably, after synthesis, the resulting feature $\widetilde{X}_t$ encapsulates interpretable knowledge encompassing aspects including but not limited to the object being gazed at, the relative spatial position of each operatable object, and the orientation of the manipulated object, etc. These synthesized features serve as a foundation for downstream models, facilitating a deeper understanding of human-robot interaction dynamics and the intricacies of interaction within teleoperation scenarios.

Additionally, it learns a time-homogeneious action transition model $\pi : \mathcal{A} \to \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}}$ represents the probability simplex on $\mathcal{A}$, and $\pi(a'|a)$ signifies the probability of taking action a' following action a. It captures the likelihood of transitioning from one action to another from labelled training data D_{train} as follows:

$$\pi(a'|a) = \frac{n_{D_{train}}(a, a')}{n_{D_{train}}(a)}, \quad \forall a, a' \in D_{train}, \tag{3}$$

where $n_{D_{train}}(a) := \sum_{t \in |D_{train}|} \mathbb{I}\{a_t = a\}$ denotes the total occurrences of action a in D_{train}, and $n_{D_{train}}(a, a') = \sum_{t \in |D_{train}|} \mathbb{I}\{a_t = a, a_{t+1} = a'\}$ indicates the number of times that action a' follows action a in D_{train}. While we consider π to encode the first-order Markovian dependencies between actions, it can readily extend to capture the higher-order dependencies when certain actions are more likely to follow specific sequences of actions.

This action transition model specifies all permissible sequences of actions in performing each task h, and is useful in refining the inference of action primitives in the subsequent low-level intention inference model. Given a sequence of actions $\{a_\tau\}_{\tau=1}^{t}$, the likelihood of this sequence occurring is computed as:

$$p(a_1, \ldots, a_t) = p(a_1) \prod_{\tau=1}^{t-1} \pi(a_{\tau+1}|a_\tau).$$

If the inferred action sequence yields zero likelihood deduced by π, a new action primitive a_t is regenerated from the last layer of the model until it results in a valid sequence. Hence, precise learning of the action transition model is crucial. We show in Theorem 1 that π can be efficiently learned without a large training dataset. The learning error of π has logarithmic dependence on both the size of training data D_{train} and the cardinality of the action space $\mathcal{A}$.

Theorem 1 *For a fixed dataset D_{train} of size $|D_{train}|$, let the true action transition model be π^*, then for some $\delta \in (0, 1)$, with probability of at least $1 - \delta$,*

$$\|\pi^*(\cdot|a) - \pi(\cdot|a)\|_1 \leq \sqrt{\frac{32 \log(2|\mathcal{A}||D_{train}|/\delta)}{\max\{1, n_{D_{train}}(a)\}}}, \quad \forall a \in \mathcal{A}. \tag{4}$$

Proof of Theorem 1. The proof directly follows from the Azuma–Hoeffding inequality with a union bound. □

In essence, the knowledge synthesizer empowers the robot with cognitive capabilities to understand the input observations and synthesize useful behavioral patterns.

B.2 Low-level Intention Inference Model

In addressing the inference of low-level action intentions, we employ an RNN architecture as our backbone model. Leveraging recurrent connections and sequential processing nature to retain a memory of previous inputs, it is well-suited for modeling temporal dependencies in sequential observations captured in each teleoperated task.

The model takes the synthesized features $\{\widetilde{X}_t\}_{t=1}^{|D_{train}|}$ generated from the knowledge synthesizer as input. To maximize the conditional log-likelihood $\log p_\theta(a_t^*|\widetilde{X}_t)$, we adopt a cross-entropy (CE) loss:

$$\mathcal{L}_{ce} = -\sum_{t=1}^{|D_{train}|} \log p_\theta(a_t^*|\widetilde{X}_t). \tag{5}$$

We use the ground truth action primitives $\{a_t^*\}_{t=1}^{|D_{train}|}$ for training and synthesized features $\{\widetilde{X}_t\}_{t=1}^{|D_{test}|}$ of D_{test} for inference.

In the context of teleoperation, different human operators may exhibit significantly different behaviors in controlling remote robots. Such variations may arise from factors such as operating positions and angles, environmental conditions, or individual traits like height and handedness. To facilitate precise inference across diverse human partners, we employ representation learning to capture the underlying patterns and structures of human behaviors in a latent embedding space. We first map the synthesized features $\{\widetilde{X}_t\}_{t=1}^{|D_{train}|}$ into a latent space in $\mathbb{R}^{d_z}$ to obtain the corresponding latent representations:

$$\widetilde{Z}_t = \phi_x(\widetilde{X}_t), \ \ \forall t \in [1, \ldots, |D_{train}|],$$

in which ϕ_x is a fully-connected layer with tanh activation.

To implement contrastive learning, we construct pairs of data $\{(\widetilde{Z}_{t_i}, \widetilde{Z}_{t_j})\}$. For each latent feature embedding $\widetilde{Z}_{t_i}$, we randomly select k embeddings sharing the same action intention $a_{t_i}^*$ as positive samples, and another k embeddings with different action intentions as negative pairs. We aim to pull the matched embeddings of the same action primitive together, while pushing those representing distinct actions apart. In essence, our model aims to maximize the mutual information (MI) between action embeddings that are matched, which describe the same semantic meaning. We thus minimize a weighted supervised InfoNCE loss Khosla et al. (2020), which serves as the lower bound of MI:

$$\mathcal{L}_{wnce} = -\sum_{i=1}^{|D_{train}|} \log \frac{\exp(s_{i,i})}{\sum\limits_{a_i^* \neq a_j^*} \exp(s_{i,j}) + \alpha \sum\limits_{a_i^* = a_j^*} \exp(s_{i,j})}, \tag{6}$$

where $s_{i,j} := sim(\widetilde{Z}_{t_i}, \widetilde{Z}_{t_j})/\tau)$, sim measures the similarity between two embeddings, τ is the temperature parameter, and α is a weighted hyperparameter. By doing so, our model preserves the semantic meanings of the observation sequences in the latent space. It maintains interpretability and explainability, allowing us to relate the latent representation to the original teleoperation data, and enhancing our understanding of how different actions manifest in the latent space.

Taken together, we define an *intention-aware loss* $\mathcal{L}_{tl}$ as a mixture of cross-entropy loss $\mathcal{L}_{ce}$ and contrastive loss $\mathcal{L}_{wnce}$:

$$\mathcal{L}_{tl} = (1 - \lambda)\mathcal{L}_{ce} + \lambda\mathcal{L}_{wnce}, \tag{7}$$

where λ is an adaptive hyperparameter being adjusted iteratively to ensure the best learning performance. Intuitively, optimizing our model with $\mathcal{L}_{tl}$ encourages learning laten human behavior patterns in the embedding space, preserving the temporal coherence, sematic interpretability, and mutual information of low-level intention.

B.3 Structure Discovery of Intention Hierarchy

The human intention hierarchy outlines the step-by-step intermediate procedures for the successful execution of complex tasks. To effectively capture the underlying intention hierarchy, we employ unsupervised hierarchical clustering to automatically discover the intermediate sub-procedures in the intention hierarchy. At each level of the intention hierarchy, sequences corresponding to the same sub-procedure are grouped into one cluster. Unlike existing model-based methods that require predefined hierarchical structures, our approach excels at automatically uncovering the latent structure of the intention hierarchy from teleoperation data, making it adaptable to various teleoperation tasks and scenarios.

The hierarchical intention model takes the latent embeddings of each testing trail from the low-level action intention model as input. Agglomerative clustering is then employed as a bottom-up approach to construct the intention hierarchy. Each embedding $\widetilde{Z}_t$ starts with its own cluster C_t. These clusters are successively merged based on the similarity measured by the Ward's linkage, which minimizes the variance of merged clusters. Essentially, it aims to merge the embeddings that constitute the same procedures for each teleoperation task, leading to clusters that represent different intermediate intentions. It effectively organizes the embedding sequences into an interpretable hierarchical structure in an unsupervised fashion, optimizing the amount of shared intention information between the observation sequences.

C Experimental Details

	Air-plane	Glider	Char-acter	Cottage	Shelf	Drag-onfly
Objects	6	8	8	5	8	5
Actions	14	19	14	8	14	12
Action Primitives	"take_B5", "place_P3", "fasten_GS", "fasten_with_SD", $\cdots$					

Table 2: Information of each building block assembly task. Examples of action primitives are shown in the last row. Common action primitives can be shared by different tasks.

C.1 Inference of Low-level Action Primitives

Baselines. To benchmark the performance of inferring low-level action intentions, we compare RoHIE against the following baseline methods: (1) Random: a model that randomly infers action primitives from those observed in training trials; (2) Majority: a model that consistently infers the most frequently occurring action primitive observed in training trials; (3) NaiveBayes: a Naive Bayes model employing Gaussian priors and Gaussian likelihoods; (4) SVM: a support vector machine model with linear kernel; (5) DecisionTree: a non-parametric decision tree model utilizing the Gini index as the splitting criterion; (6) RandomForest: an ensemble model comprising 50 decision trees with the Gini index as the splitting criterion; (7) MLP: a multilayer perception model (8) RNN_baseline: an RNN baseline model that directly takes the raw collected data as input.

Training Details. For a fair comparison, all neural network baselines are equipped with identical layer structures and trained using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 256, and training epochs set to 200. The optimal values of all hyperparameters are determined through grid search. For baselines involving randomness, we conduct experiments 10 times and report the average accuracy, precision, recall, and $F1$-score in Table 1. In the case of RoHIE and RNN_baseline, the sequence length of the input is set to 60.

Inference of Low-level Action Primitives. As shown in Table 1 (left), simple baselines such as randomly inferring action primitives or selecting the most frequent action yield notably poor performance. While classical machine learning baselines and deep neural network baselines show similar performance, RoHIE consistently outperforms them across all evaluation metrics. Additionally, the inference accuracy of each action primitive is visualized in Figure 7 (left), affirming RoHIE's capability to achieve satisfactory inference results for low-level action intentions.

C.2 Generalization and Robustness

In practice, human operators frequently exhibit varied behaviors during teleoperation. As an illustration, Figure 4 depicts several different scenarios in which human participants demonstrate substantially different behaviors when teleoperating robot arms. To evaluate RoHIE's generalization ability, we conduct two sets of experiments to assess (1) its capability to accurately infer the intentions of new human partners who have not been involved in the training process, and (2) its robustness to maintain effective inference capabilities when training data is limited.

Genearlization to New Human Partners. We deploy RoHIE to infer the intention of a new human partner during the execution of assembly tasks. Figure 5 shows the inference results at each time step when the operator undertakes the "airplane" and "shelf" assembly tasks, respectively. While occasional inference errors may arise, RoHIE consistently avoids producing irrational action primitives unrelated to the ongoing assembly task. Notably, the inference errors typically occur during transitions between completed actions and the initiation of subsequent actions. Hence, RoHIE demonstrates exceptional generalization capabilities in accurately inferring the intentions of new human partners.

Robustness of Inference Performance. We proceed to evaluate the robustness of RoHIE by varying the number of trails used for training. Specifically, for human participants selected for testing, all of their trials will be excluded from the training. As shown in Table 1 (right), our method maintains high inference accuracy even with a small number of training trials, whereas the performance of the Random Forest algorithm and RNN baseline rapidly deteriorates as the size of the training dataset

decreases. Furthermore, as demonstrated in Figure 6, compared to the RNN baseline, RoHIE effectively pulls the sequences of the same action together while separating sequences corresponding to different action primitives. By optimizing RoHIE with the proposed intention-aware loss using representation learning, it learns compact semantic representations of low-level actions in the latent space, capturing common structural information from diverse human operators, thereby endowing it with robust inference capabilities. Interestingly, RoHIE achieves slightly improved inference accuracy with a train-test split of $50\% - 50\%$, compared to scenarios with larger training datasets. This phenomenon suggests that while RoHIE provides robust inference by extracting semantic commonalities of human behavior in the latent space, biases and noise present in training observations may still exert some influence on the final performance, albeit to a lesser extent.

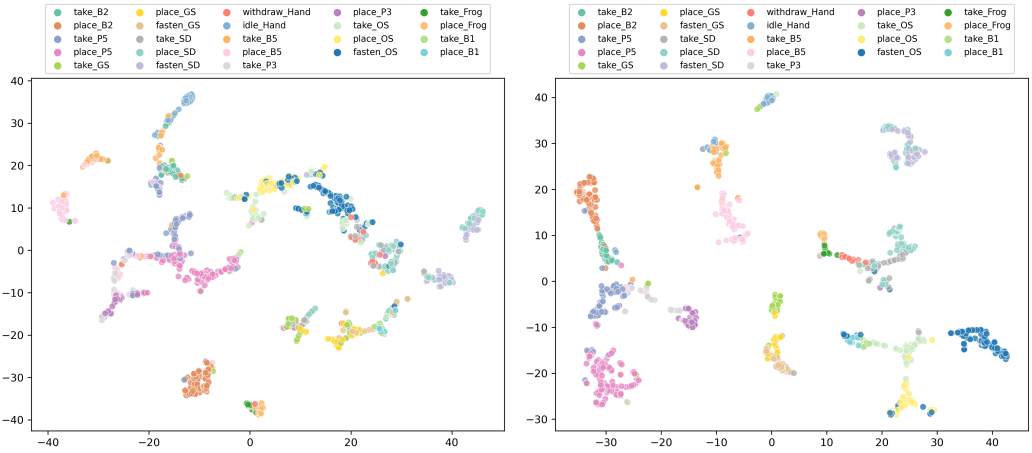

Figure 6: Left: t-SNE of action embeddings of RNN baseline. Right: t-SNE of action embeddings of RoHIE. RoHIE generates semantically-better representations for low-level action primitives in the embedding space by pulling the sequences of the same action intention together while pushing the distinct ones away.

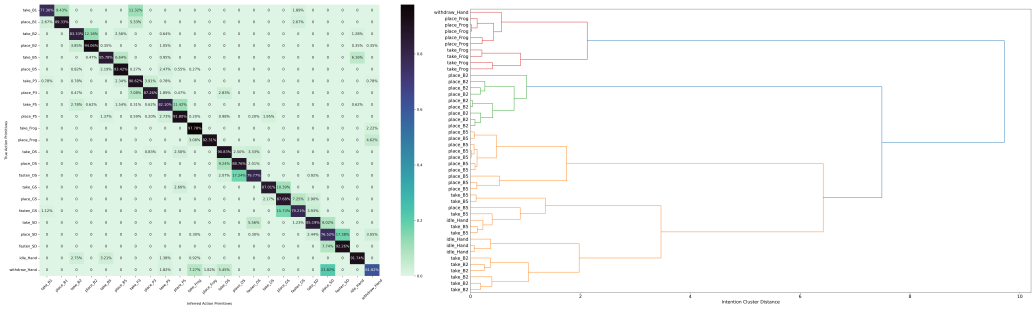

Figure 7: Left: Inference accuracy of each low-level action primitive. Right: inferred intention hierarchy for the assembly task "Cottage". Each action primitive a_t on the left hand side represents a latent embedding sequence corresponding to a_t.

C.3 Discovery of Intention Hierarchy

To discover the intention hierarchy for each task, we build a hierarchical tree (a.k.a. dendrogram) to depict the hierarchical relationship among different intention levels, thereby grouping latent embeddings with similar low-level intentions to form higher-level task intention. One of the resulting inferred intention hierarchies is shown in Figure 7 (right). We analyze the dendrogram to identify natural breakpoints to determine different levels of the intention hierarchy. In the building block assembly dataset, RoHIE successfully retrieves the intention hierarchy for each task, which consists of intermediate sub-procedures to help the remote robot understand and discover the workflows to complete each high-level assembly task. We remark that such automatically-discovered intention hierarchies are both semantically meaningful, and can be further incorporated with probabilistic models such as hierarchical hidden Markov models (HHMMs).

D Future Directions

Here, we highlight several interesting directions for future exploration. In light of the recent advances in large language models (LLMs), it is beneficial to employ the power of LLMs to further automate the knowledge synthesis of RoHIE. Besides, enabling uncertainty quantification of RoHIE by further incorporating intention hierarchies with probabilistic models can be imperative when performing complex tasks with shared sub-procedures. In summary, RoHIE is an effective and comprehensive solution for intention estimation in human-robot teleoperation.