

TASC: Task-Aware Shared Control for Teleoperated Manipulation

Ze Fu¹, Pinhao Song¹, Yutong Hu¹, Renaud Detry^{1,2}

Abstract—We present TASC, a **Task-Aware Shared Control** framework for teleoperated manipulation that infers task-level user intent and provides assistance throughout the task. To support everyday tasks without predefined knowledge, TASC constructs an open-vocabulary interaction graph from visual input to represent functional object relationships, and infers user intent accordingly. A shared control policy then provides rotation assistance during both grasping and object interaction, guided by spatial constraints predicted by a vision-language model. Our method addresses two key challenges in general-purpose, long-horizon shared control: (1) understanding and inferring task-level user intent, and (2) generalizing assistance across diverse objects and tasks. Experiments in both simulation and the real world demonstrate that TASC improves task efficiency and reduces user input effort compared to prior methods. To the best of our knowledge, this is the first shared control framework that supports everyday manipulation tasks with zero-shot generalization. The code that supports our experiments is publicly available at <https://github.com/fitz0401/tasc>.

I. INTRODUCTION

Teleoperation enables human operators to control robots in unstructured or safety-critical environments, where full autonomy remains unreliable due to perceptual uncertainty (e.g., underwater exploration [1]), task ambiguity (e.g., assistive robotics [2]), or the need for human judgment (e.g., surgical robotics [3]). Although easy to deploy, direct operation of high-degree-of-freedom (DoF) robotic arms remains cognitively demanding and physically fatiguing, especially for novice operators. To reduce this burden, shared control frameworks blend user input with autonomous assistance, leveraging sensor feedback to filter user input noise and improve precision [4], [5].

A shared control system assists by interpreting the user’s intent, which can be expressed through various modalities: language is well suited for specifying high-level task goals [6], eye gaze can indicate object-level interest [7], and hand-operated devices (e.g., joysticks and keyboards) are effective for conveying motion-level preferences [8]. In this work, we focus on intent expressed exclusively through hand-operated input. This choice is motivated by the fact that motion control is a natural way for humans to execute tasks, offering strong procedural compatibility and precise control over robot behavior. We believe that shared control should be fundamentally driven by motion-based input, with high-level modalities such as language serving as optional

complements once intent inference from low-level signals is well established.

Existing shared control methods based on motion input remain far from providing general-purpose, long-horizon assistance in everyday environments [9]. We identify two key challenges that hinder progress:

(1) Task understanding and inference. Inferring user intent from motion commands is difficult in multi-step everyday tasks. For example, pouring water involves grasping a kettle, moving it over a cup, and executing the pour. To support such tasks, shared control systems must identify the current task and recognize task progression. **(2) Task generalization.** A general-purpose shared control system should assist in diverse tasks involving varied objects. Prior works are largely restricted to well-structured problems, such as target reaching [8], [10], [11] and grasping [12], [6], which struggle to generalize to everyday manipulation involving diverse object-object interactions and high-DoF continuous action spaces.

Recent advances in vision-language models (VLMs) offer a promising direction for general-purpose, long-horizon shared control. Trained on large-scale multimodal data, VLMs encode rich commonsense knowledge about object semantics and interactions. They can infer functional relationships between objects [13], [14], and reason about plausible tasks from a scene [15], [16].

Based on this insight, we propose TASC (see Fig. 1), a task-aware shared control framework that supports task-level intent inference and assistance throughout task execution. Our system consists of two key components:

(1) First, we construct an **open-vocabulary interaction graph** capturing possible object-object relationships and corresponding interaction constraints through VLMs. This graph representation helps our system to understand the composition and progression of multi-step tasks. Unlike prior approaches that focus only on short-term goals, our method enables inference of operator task-level intent and supports assistance over extended tasks.

(2) Second, we present a **multi-step shared control policy** composed of the grasping stage and the subsequent object interaction stage, providing a unified framework for diverse manipulation tasks. In the grasping stage, the system assists the user by selecting an optimal grasp pose using an off-the-shelf grasp planner. In the object interaction stage, it infers spatial affordances between the grasped object and its target via the interaction graph, and formulates an optimization problem to solve for a suitable pose for interaction. This two-stage design reflects common tool-use patterns and enables generalization to diverse manipulation tasks without predefined task descriptions or demonstrations.

¹KU Leuven, Dept. Mechanical Engineering, Research unit Robotics, Automation and Mechatronics, B-3000 Leuven, Belgium. firstname.lastname@kuleuven.be

²KU Leuven, Dept. Electrical Engineering, Research unit Processing Speech and Images, B-3000 Leuven, Belgium.

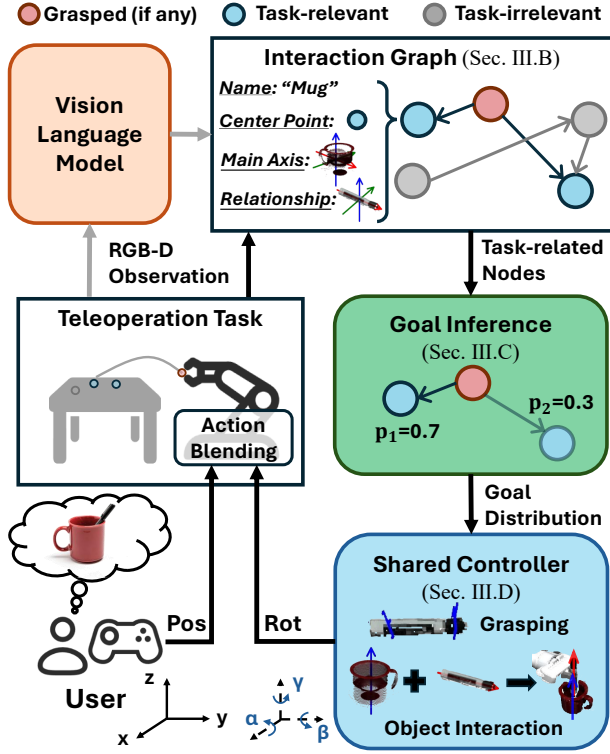


Fig. 1: Overview of our TASC framework. The teleoperation loop begins with RGB-D observations of the scene, which are processed jointly with a VLM to construct an interaction graph (see Sec. III-B) encoding object centers, main axes, and functional relationships. In the interaction graph, the red node denotes the grasped object (if any), blue nodes denote task-relevant objects that serve either as grasp targets or as functionally related objects, and gray nodes denote task-irrelevant objects. The inference module (see Sec. III-C) fuses this graph with motion cues to estimate a goal distribution. Guided by this distribution, the shared controller (see Sec. III-D) blends the user’s translational input (Pos) with assistive rotations (Rot): during grasping, it assists with a grasp planner, while during object interaction it enforces alignment of the two objects’ main axes.

Our experiments, with both simulated and real-world teleoperated manipulation, show that TASC allows users to complete everyday tasks with reduced input burden and improved efficiency, and generalizes across multiple manipulation tasks and object configurations without task-specific training.

The main contributions of this paper are:

- A task-aware shared control framework that infers user intent from historical control behavior and scene semantics, represented via an open-vocabulary interaction graph. This enables inference of multi-step goals without predefined task descriptions.
- A shared control policy that provides rotation assistance throughout the task: selecting optimal grasp poses during the grasping stage, and solving optimization problems to align objects according to inferred spatial affordances during the object interaction stage.

- Experimental validations on multi-step teleoperation tasks, demonstrating improved task efficiency and reduced user input compared to pure teleoperation and shared control baselines.

II. RELATED WORK

Shared Control for high-DoF robotic arms has been widely explored to reduce user burden during teleoperation. Early works primarily focus on assisting EEF translation. For example, Gottardi et al. [17] propose an artificial potential field augmented with dynamically generated escape points to guide the robot toward likely goals while avoiding obstacles. Beyond planning-based methods, learning-based approaches leverage collected user data to provide predictive assistance. Robot Trajectron [8] predicts the future EEF trajectory from recent motion dynamics, enabling proactive support for reaching tasks. Yoneda et al. [18] learn assistive behaviors from demonstrations using diffusion models. However, controlling rotational DoF remains the most unintuitive and effortful aspect of teleoperation. Existing studies address this by providing predefined EEF goal poses [10], [6], [19], [20] or discrete action templates [21]. While effective in structured settings, these approaches struggle to generalize to everyday manipulation tasks. Another line of work mitigates this challenge by introducing mode-switching strategies between translation and rotation control [22], [23], allowing efficient assistance under motion-based user input, but still placing the burden of precise orientation control on the operator.

In addition to limited support for rotation, most existing methods also restrict the task scope to short-horizon behaviors such as reaching [8], [10], [11], grasping [12], [6], or obstacle avoidance [17], [24]. This narrow focus limits their applicability in real-world scenarios. Providing task-level assistance requires richer contextual information beyond user input history. Recent work such as VOSA [25] takes a step in this direction by leveraging vision-based object detection to infer user intent in a zero-shot manner. However, it still offers limited support for post-grasp object interactions and underutilizes semantic cues available from modern foundation models.

Graph-based Representations are well-suited for modeling object relationships. In indoor navigation, scene graphs have been used to represent spatial layouts and semantic relationships to support language-guided path planning [16]. In manipulation, action-conditioned scene graphs [15] have enabled interactive exploration and planning over long-horizon tasks by capturing how object states and relationships evolve in response to robot actions. Compared to prior works [16], [15], [26], [27] that primarily model spatial relations (e.g., on, inside), we focus on capturing potential functional interactions between object pairs. In this direction, prior efforts [28], [29] in object-object interactions (OOI) have proposed taxonomies and datasets to classify spatio-temporal relationships between objects. However, their applications in robotic manipulation remain underexplored.

Affordance for Manipulation provides actionable priors that support tool usage and motion planning. Early works [30], [31], [32] focus on learning single-object affordances through data collection and supervised training, while recent efforts [33], [34] extend to modeling object-object affordances. However, these approaches are constrained by data collection efficiency and limited generalization. The emergence of foundation models has made zero-shot affordance inference increasingly feasible because of their ability for semantic scene understanding and commonsense reasoning. Recent works [35], [36], [13] use VLMs to extract keypoint constraints between objects under task instructions. However, due to the limited spatial grounding capabilities of current VLMs, the resulting keypoints can be unstable or inconsistent across queries. Inspired by OmniManip [14], we instead extract task-relevant pose alignment constraints between object pairs via VLMs. This provides a stable and concise representation that is well-suited for assisting rotation in shared control.

III. METHOD

This section first formalizes the problem of multi-step teleoperation under shared control. To provide general-purpose, long-horizon assistance, we introduce TASC (see Fig. 1), a framework consisting of: (1) an open-vocabulary interaction graph that encodes functional object relationships (Sec. III-B); (2) a goal inference module that maintains a belief over the user’s task-level intent (Sec. III-C); and (3) a shared control policy that provides rotation assistance during both grasping and post-grasp object interaction (Sec. III-D).

A. Problem Formulation

We study the problem of shared control teleoperation for multi-step manipulation tasks. The user operates the robot via a hand-operated input device. We assume a finite task space $\mathcal{T}$ defined over N objects, where each task involves functionally meaningful interaction between objects in the scene (e.g., placing a banana on a plate). The user intends to complete a task in $\mathcal{T}$ and this intent is unknown to the system. To reduce user burden, our aim is to allow the user to control only the translational DoF of the EEF, with the system providing rotation assistance as needed. We seek to achieve this in a zero-shot manner, representing an essential step toward practical, everyday use of shared control.

During teleoperation, the user issues a sequence of control inputs $u_h \in \mathbb{R}^3$, where $u_h = [x, y, z]^T$ denotes the desired translational motion of the EEF. Our objective is to infer the user’s intended task within the unknown task space $\mathcal{T}$ by leveraging both u_h and affordance relationships between objects. The system then generates assistive rotational commands $u_r = [\alpha, \beta, \gamma]^T$ to support effective execution.

B. Open-vocabulary Interaction Graph

To support task-level intent inference, we construct an *open-vocabulary interaction graph* that encodes possible object-object functional relationships, which is shown in

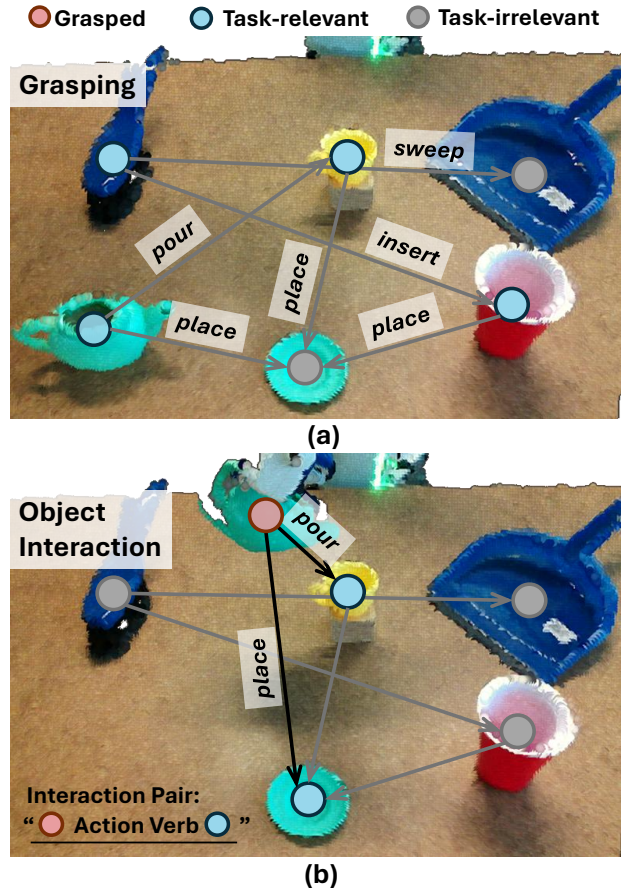


Fig. 2: The interaction graph with nodes in different activation states: (a) During grasping, nodes with outgoing edges (i.e., functional tools) are marked active; (b) during object interaction, only the neighbors of the grasped node are considered.

Fig. 2. We define an interaction as a functional composition between two objects that supports a meaningful task, represented as a triplet: *object_A – action verb – object_B* (e.g., “hammer-hit-nail”).

Given an RGB image $\mathcal{R}$ capturing the workspace objects $\mathcal{O} = \{o_i\}_{i=1}^N$, we query a VLM to extract a set of object names and candidate interaction triplets. Unlike typical object recognition pipelines [37], [38] that attempt to detect all objects and categorize them, we prompt the model to identify only objects that are functionally relevant, which avoids the need to tune recognition granularity manually. We then construct a directed graph $\mathcal{G} = (V, E)$, where each node $v_i \in V$ stores the spatial and geometric information of an object $o_i \in \mathcal{O}$, and each edge $e_{ij} \in E$ represents an interaction from o_i to o_j via an action verb.

To ground the graph in the physical world, we apply GroundedSAM [39] using the object names to segment masks $\mathcal{M}$. For each mask $m_j \in \mathcal{M}$, we compute its center and project it into the world coordinate frame $\mathbb{R}^3$ via a calibrated RGB-D camera to obtain a 3D position p_j . We also extract a partial point cloud pcd_j for each object. The

Object Interaction Assistance is activated after a successful grasp. Manipulation tasks involving functional object interactions usually require specific pose relationships

between objects. These requirements can be modeled as desired alignments between their local coordinate frames. To enable such assistance, we identify the interaction pair from $\mathcal{G}$: the grasped object o_A and its associated target object o_B , as shown in Fig. 3. For each object, we compute its canonical axes (X, Y, Z) using the oriented bounding box from its point cloud. We then render axis-overlaid object images and prompt a VLM to identify which axes should be aligned or anti-aligned. This produces a set of constraints $\mathcal{C} = \{c_i = (a_i, b_i, s_i) \mid i = 1, \dots, k\}$, where each c_i specifies that axis a_i of o_A should be aligned (if $s_i = 1$) or anti-aligned (if $s_i = -1$) with axis b_i of o_B , and k is the number of predicted constraints. Given the estimated poses of o_A and o_B and the corresponding $\mathcal{C}$, we formulate an optimization problem to compute the minimal rotation $R_g \in SO(3)$ that best satisfies the desired axis alignments:

$$\min_{R_g \in SO(3)} \sum_{c_i \in \mathcal{C}} \|R_g \cdot a_i - s_i \cdot b_i\|^2 \quad (5)$$

During shared control, this predicted rotation is blended with the user’s translational input to support precise alignment. This module extends shared control beyond grasping, enabling task-aware assistance for post-grasp object interaction. We validate its effectiveness in multi-step manipulation tasks such as placing and hammering, both in simulation and on real hardware.

IV. EXPERIMENTS

To comprehensively evaluate the performance of TASC, we conducted both simulation studies and real-world shared control experiments. All experiments were performed on the Franka Research 3 robotic platform. For our method, we used GPT-4o [42] to extract task-relevant object relations for constructing the interaction graph. Prompts used for querying the model are provided in our code repository.

A. Experiment in the Simulation

We evaluate our method in a tabletop teleoperation environment to answer three main questions: (1) Can TASC achieve accurate task-level intent inference? (2) Can TASC improve efficiency and reduce input effort? (3) Can TASC enhance user experience in teleoperated manipulation?

Simulation Setup. As shown in Fig. 4, we built a tabletop manipulation environment using Robosuite v1.5 [43] with six objects: banana, plate, marker, mug, hammer, and a wooden block with a peg. All object models are from the YCB dataset [44], except that the wooden block is customized by attaching a cylindrical mesh to simulate the peg. We defined three multi-step manipulation tasks of increasing difficulty:

- 1) **Place:** Place the banana onto the plate. Considered successful if the banana’s center lies within the plate region. This is a simple pick-and-place task with minimal orientation requirements.
- 2) **Insert:** Insert the marker into the mug. Requires the pen to be grasped and placed nearly upright inside the mug, demanding accurate orientation alignment.

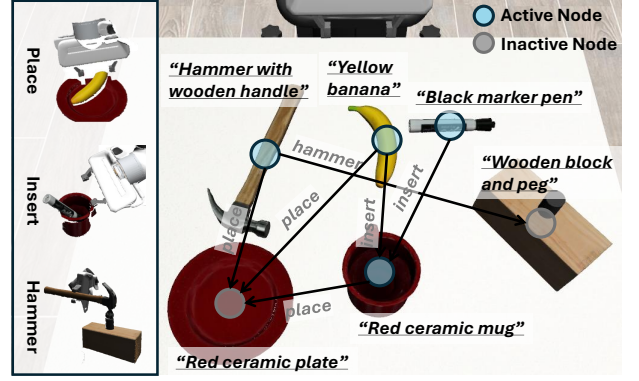


Fig. 4: Simulation setup and initial states of the objects for multi-step shared control tasks.

- 3) **Hammer:** Hit the peg with the hammer. Success requires the hammerhead to contact the peg head with an orientation error below 10° , simulating a precise downward strike.

All tasks involve both grasping and post-grasp object interaction. The user and the system both observe the environment from a single calibrated top-down RGB-D view. Notably, our system has no access to any prior knowledge about object properties or task details.

Protocol. The environment provides a teleoperation interface enabling users to control the EEF position, orientation, and the gripper state via keyboard. Each trial starts with a fixed initial EEF pose, while object poses are randomized. We recruited 10 novice participants from the local community. Each received brief training to operate the robot using a keyboard. During the formal experiment, participants performed the three tasks in a randomized order, and the order of shared control methods was also randomized.

Baselines and Metrics. We compared TASC with three baselines.

- **Teleop.** Pure user control without any assistance.
- **TASC⁻.** An ablation of our method without using the interaction graph. In the grasping stage, any object with at least one valid grasp predicted by the grasp planner is treated as a candidate goal. In the object interaction stage, all remaining objects on the table (except the grasped one) are included in the candidate goal set, regardless of functionality.
- **VOSA [25].** A vision-only, zero-shot shared autonomy framework that supports user intent recognition and pick-and-place assistance. It assists only with position control and assumes a predefined set of possible placement locations. To minimize comparison bias, we implement the same perception module (GroundedSAM [39]) for VOSA as in TASC.

For each trial, we recorded the completion time, trajectory length of the EEF, and the number of control inputs (button presses) during the task execution. In addition to these objective indicators, we also collected four subjective indicators through a questionnaire, each scored on a scale from 0 to

TABLE I: Performance across teleoperation tasks. Each cell shows (Time (s) / Trajectory Length (m) / Inputs).

Method	Banana-Plate	Marker-Mug	Hammer-Peg
Teleop	54.7 / 1.50 / 135.4	99.9 / 2.16 / 228.3	112.9 / 3.40 / 338.6
VOSA	61.9 / 1.57 / 136.9	97.8 / 2.41 / 222.0	132.9 / 3.80 / 341.2
TASC ⁻	91.0 / 2.26 / 178.7	124.5 / 2.99 / 267.9	73.9 / 2.42 / 218.0
TASC	66.4 / 1.71 / 116.6	54.7 / 1.52 / 95.6	72.1 / 2.42 / 186.7

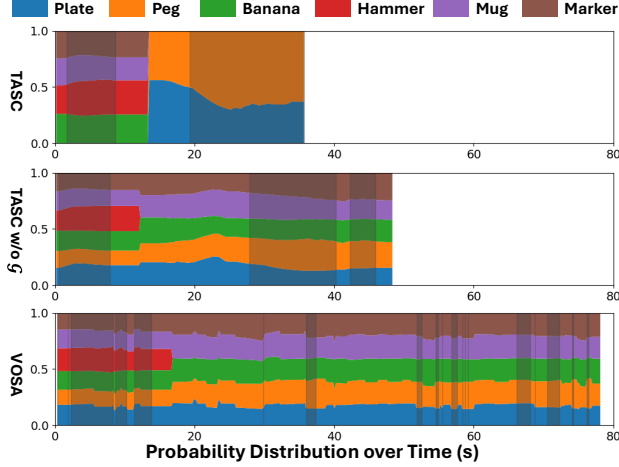


Fig. 5: Goal probability distribution over time across three shared control methods; shaded regions denote correct goal predictions during the grasping and object interaction phases.

10:

- Control: I felt in control of the robot.
- Expectation: The robot did what I wanted.
- Usability: If I was going to teleoperate a robotic arm, I would like to use this system.
- Satisfaction: I was satisfied with my performance.

Results and Analysis. Results are shown in Table I. TASC consistently achieves fewer control inputs across all tasks, indicating lower user burden. In the simple *banana-plate* task which primarily involves straightforward reaching, the advantage of our method is less evident; however, TASC still reduces user inputs compared to pure teleoperation. For the more challenging *marker-mug* and *hammer-peg* tasks involving precise alignment, TASC achieves faster task execution efficiency and leads to shorter trajectories. Compared to VOSA, which lacks reasoning about object-object interactions, TASC leverages spatial affordances to alleviate the most fatiguing rotational controls and accelerate task progression. In contrast to its ablated version, TASC benefits from semantic reasoning to guide task-level intent inference, enabling faster and intuitive teleoperation.

We also visualize the evolution of the predicted goal distribution $p(g|\xi^{0:t})$ in Fig. 5, using a representative hammer-hit-peg task. Each colored area represents the probability assigned to a candidate goal, while shaded regions indicate periods in which the system correctly identifies the intended object during the grasping (hammer) and object interaction (peg) stages. Our method exhibits early convergence to the

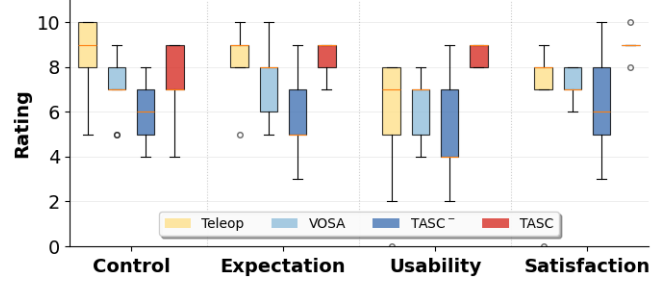


Fig. 6: User study results on control, expectation, usability, and satisfaction.

correct goal and maintains stable predictions throughout. In contrast, baseline methods (TASC⁻ and VOSA) show delayed convergence and fluctuating predictions, particularly during the object interaction stage. This indicates that our intent inference mechanism not only improves accuracy, but also enhances the temporal consistency of assistance.

As illustrated in Fig. 6, TASC achieves consistently higher ratings on most subjective metrics compared to the baselines. While pure teleoperation naturally offers the highest sense of control, TASC maintains a comparable level in this metric and improves how well the robot behaves as expected, demonstrating a favorable balance between user control and assistance. Furthermore, TASC outperforms all baselines in both usability and satisfaction, confirming its effectiveness in enhancing the overall teleoperation experience.

B. Experiment in the Real-World

In this real-world experiment, we focus on evaluating the generalization capability of TASC, i.e., whether it can handle unseen real-world scenes without any task-specific tuning.

Settings. The experiments were conducted in a tabletop workspace as illustrated in Fig. 7, with a top-down view provided by an Intel RealSense L515 camera mounted in front of the robot. We prepared diverse everyday manipulation tasks with varying object interactions and configurations, while both objects and task details were unknown to TASC. The user controlled the robot EEF translation, rotation, and gripper state using a joystick, with the left side of the joystick assigned to rotation and the right side to translation. A TASC-assisted task is considered successful if the user can complete the task using only the right side of the joystick.

Results. As illustrated in Fig. 8, TASC demonstrates strong generalization across a wide range of real-world tasks, including placing, aligning, and stacking. Despite significant variations in object geometry and interaction patterns, TASC successfully assists users in completing these tasks without any task-specific tuning. Compared to pure teleoperation, users can complete tasks with one hand by controlling only simple translational motions.

We further evaluate TASC under variations in initial object placements and multi-object interaction scenarios. As shown in Fig. 9(a), three representative tasks are demonstrated: sweeping with a brush, inserting a roller into a holder, and pouring with a kettle. TASC generalizes to different object

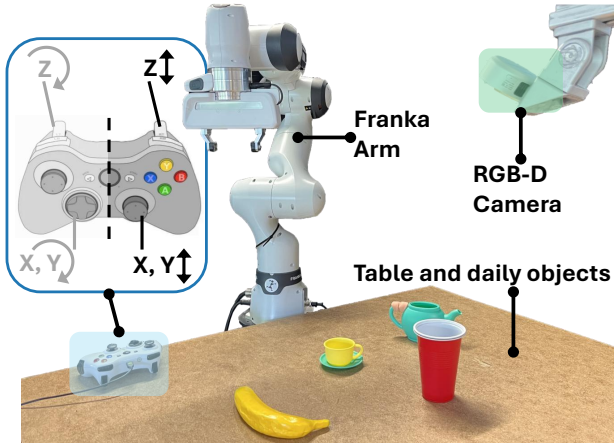


Fig. 7: Real-world experimental setup with a Franka Research 3 arm, tabletop objects, and a top-down RGB-D camera.

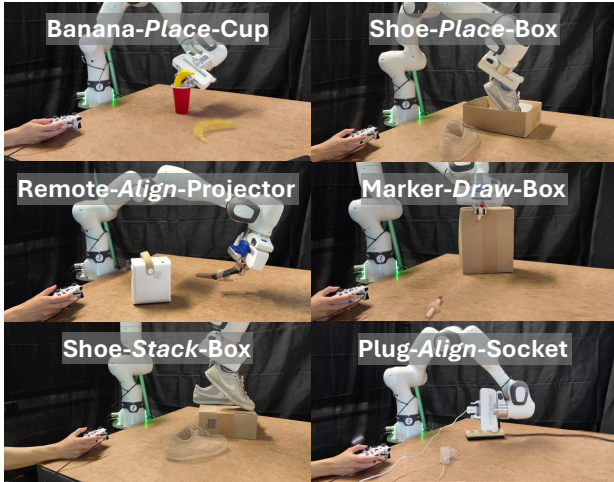


Fig. 8: Snapshots of successful shared-control task executions with TASC, where semi-transparent objects indicate initial positions.

configurations without task-specific training. In Fig. 9(b), both the remote controller and the kettle interact sequentially with multiple objects. TASC correctly predicts the intended interaction pairs and adapts its assistance accordingly, thereby enabling users to accomplish long-horizon manipulation tasks.

V. CONCLUSIONS

We presented a task-aware shared control framework for everyday multi-step manipulation. Our key insight is to leverage foundation models to extract task-relevant object relations and spatial affordances, which are structured into an open-vocabulary interaction graph. This enables efficient goal inference and assistance during both grasping and object interaction. We demonstrated that our system improves task execution efficiency and reduces user input effort compared to existing shared control baselines. One current limitation is that our stage decomposition may not generalize well to

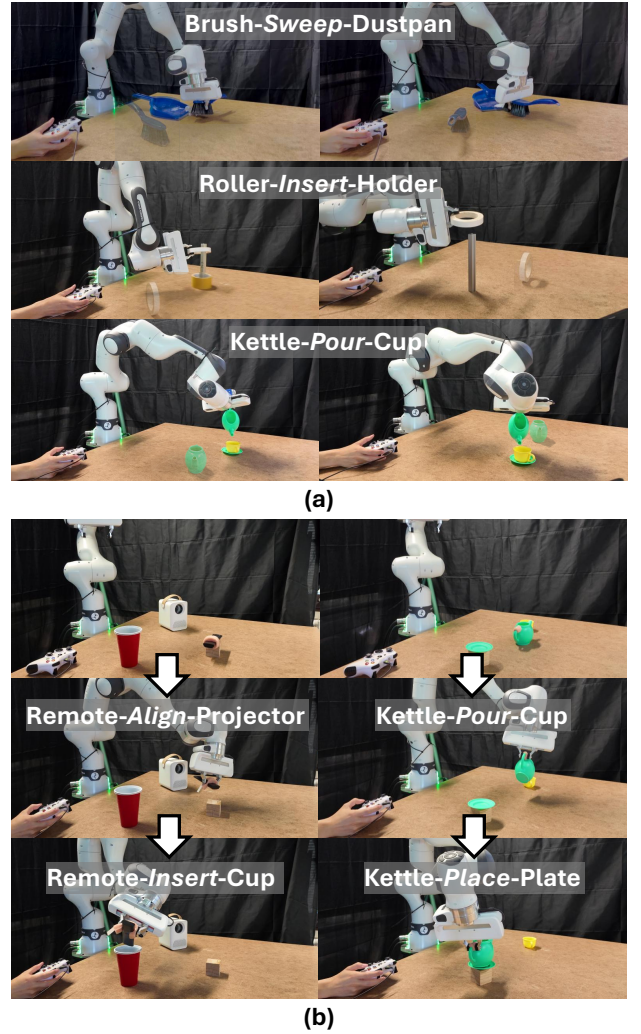


Fig. 9: Real-world shared-control experiments with TASC, where semi-transparent objects indicate initial positions: (a) variations in object configurations; (b) multi-object, multi-step manipulations.

tasks involving articulated objects, such as opening drawers. In future work, we plan to extend our framework to more complex tool-use scenarios, explore integration with other input modalities such as language, and study the applicability of our framework to BCI-controlled manipulation.

REFERENCES

- [1] Q. Lin and C. Kuo, "Virtual tele-operation of underwater robots," in *Proceedings of International Conference on Robotics and Automation*, vol. 2. IEEE, 1997, pp. 1022–1027.
- [2] Y. Liu, W. Su, Z. Li, G. Shi, X. Chu, Y. Kang, and W. Shang, "Motor-imagery-based teleoperation of a dual-arm robot performing manipulation tasks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 3, pp. 414–424, 2018.
- [3] G. H. Ballantyne and F. Moll, "The da vinci telerobotic surgical system: the virtual operative field and telepresence surgery," *Surgical Clinics*, vol. 83, no. 6, pp. 1293–1304, 2003.
- [4] Y. Xu, H. Zhang, L. Cao, X. Shu, and D. Zhang, "A shared control strategy for reach and grasp of multiple objects using robot vision and noninvasive brain-computer interface," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 1, pp. 360–372, 2022.

- [5] G. Gillini, P. D. Lillo, and F. Arrichiello, "An assistive shared control architecture for a robotic arm using eeg-based bci with motor imagery," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4132–4137.
- [6] J.-A. Yow, N. P. Garg, and W. T. Ang, "Shared autonomy of a robotic manipulator for grasping under human intent uncertainty using pomdps," *IEEE Transactions on Robotics*, vol. 40, pp. 332–350, 2024.
- [7] H. Admoni and S. S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," in *AAAI fall symposia*, 2016, pp. 298–303.
- [8] P. Song, P. Li, E. Aertbeliën, and R. Detry, "Robot trajectory: Trajectory prediction-based shared control for robot manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5585–5591.
- [9] G. Li, Q. Li, C. Yang, Y. Su, Z. Yuan, and X. Wu, "The classification and new trends of shared control strategies in telerobotic systems: A survey," *IEEE Transactions on Haptics*, vol. 16, no. 2, pp. 118–133, 2023.
- [10] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 2018.
- [11] D. Gopinath, S. Jain, and B. D. Argall, "Human-in-the-loop optimization of shared autonomy in assistive robotics," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 247–254, 2017.
- [12] Y. Xu, H. Zhang, L. Cao, X. Shu, and D. Zhang, "A shared control strategy for reach and grasp of multiple objects using robot vision and noninvasive brain-computer interface," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 1, pp. 360–372, 2022.
- [13] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 4573–4602.
- [14] M. Pan, J. Zhang, T. Wu, Y. Zhao, W. Gao, and H. Dong, "Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints," *arXiv preprint arXiv:2501.03841*, 2025.
- [15] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation," *arXiv preprint arXiv:2402.15487*, 2024.
- [16] Z. Yan, S. Li, Z. Wang, L. Wu, H. Wang, J. Zhu, L. Chen, and J. Liu, "Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation," *IEEE Robotics and Automation Letters*, 2025.
- [17] A. Gottardi, S. Tortora, E. Tosello, and E. Menegatti, "Shared control in robot teleoperation with improved potential fields," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 410–422, 2022.
- [18] T. Yoneda, L. Sun, B. Stadie, M. Walter *et al.*, "To the noise and back: Diffusion for shared autonomy," *arXiv preprint arXiv:2302.12244*, 2023.
- [19] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.
- [20] S. Nikolaidis, Y. X. Zhu, D. Hsu, and S. Srinivasa, "Human-robot mutual adaptation in shared autonomy," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 294–302.
- [21] G. Quere, A. Hagengruber, M. Iskandar, S. Bustamante, D. Leidner, F. Stulp, and J. Vogel, "Shared control templates for assistive robotics," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1956–1962.
- [22] L. V. Herlant, R. M. Holladay, and S. S. Srinivasa, "Assistive teleoperation of robot arms via automatic time-optimal mode switching," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 35–42.
- [23] Y. Tao, J. Yang, D. Ding, and Z. Erickson, "Lams: Llm-driven automatic mode switching for assistive teleoperation," in *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '25. IEEE Press, 2025, p. 242–251.
- [24] N. Mehr, R. Horowitz, and A. D. Dragan, "Inferring and assisting with constraints in shared autonomy," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2016, pp. 6689–6696.
- [25] A. Belsare, Z. Karimi, C. Mattson, and D. S. Brown, "Toward zero-shot user intent recognition in shared autonomy," in *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '25. IEEE Press, 2025, p. 222–231.
- [26] H. Shang, H. Wu, G. Zhai, B. Sun, F. Wang, F. Tombari, and M. Pollefeys, "Sg-tailor: Inter-object commonsense relationship reasoning for scene graph manipulation," *arXiv preprint arXiv:2503.18988*, 2025.
- [27] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu, "Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6541–6548.
- [28] A. Unmesh, R. Jain, J. Shi, V. C. Manam, H.-G. Chi, S. Chidambaram, A. Quinn, and K. Ramani, "Interacting objects: A dataset of object-object interactions for richer dynamic scene representations," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 451–458, 2023.
- [29] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, "O2o-afford: Annotation-free large-scale object-object affordance learning," in *Conference on robot learning*. PMLR, 2022, pp. 1666–1677.
- [30] F.-J. Chu, R. Xu, and P. A. Vela, "Learning affordance segmentation for real-world robotic manipulation via synthetic images," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1140–1147, 2019.
- [31] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.
- [32] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An affordance keypoint detection network for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.
- [33] H. Huang, K. Schmeckpeper, D. Wang, O. Biza, Y. Qian, H. Liu, M. Jia, R. Platt, and R. Walters, "Imagination policy: Using generative point cloud models for learning manipulation policies," *arXiv preprint arXiv:2406.11740*, 2024.
- [34] Y. Qi, Y. Ju, T. Wei, C. Chu, L. L. Wong, and H. Xu, "Two by two: Learning multi-task pairwise objects assembly for generalizable robot manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17383–17393.
- [35] F. Liu, K. Fang, P. Abbeel, and S. Levine, "Moka: Open-vocabulary robotic manipulation through mark-based visual prompting," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.
- [36] H. Huang, F. Lin, Y. Hu, S. Wang, and Y. Gao, "Copa: General robotic manipulation through spatial constraints of parts with foundation models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9488–9495.
- [37] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu *et al.*, "Recognize anything: A strong image tagging model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1724–1732.
- [38] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, C. Li, J. Yang, L. Zhang, and J. Gao, "Segment and recognize anything at any granularity," in *European Conference on Computer Vision*. Springer, 2024, pp. 467–484.
- [39] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.
- [40] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning," in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.
- [41] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics (T-RO)*, 2023.
- [42] OpenAI, "Gpt-4o technical report," 2024, accessed: 2024-06-25. [Online]. Available: <https://openai.com/index/gpt-4o>
- [43] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, Y. Zhu, and K. Lin, "robosuite: A modular simulation framework and benchmark for robot learning," in *arXiv preprint arXiv:2009.12293*, 2020.
- [44] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517.