Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Embodiment Meets Environment: Toward Context-Aware, Safe Physical Caregiving Robots

Zhanxin Wu, Ruofei Tong, Jiaying Fang, Tapomayukh Bhattacharjee

Cornell University

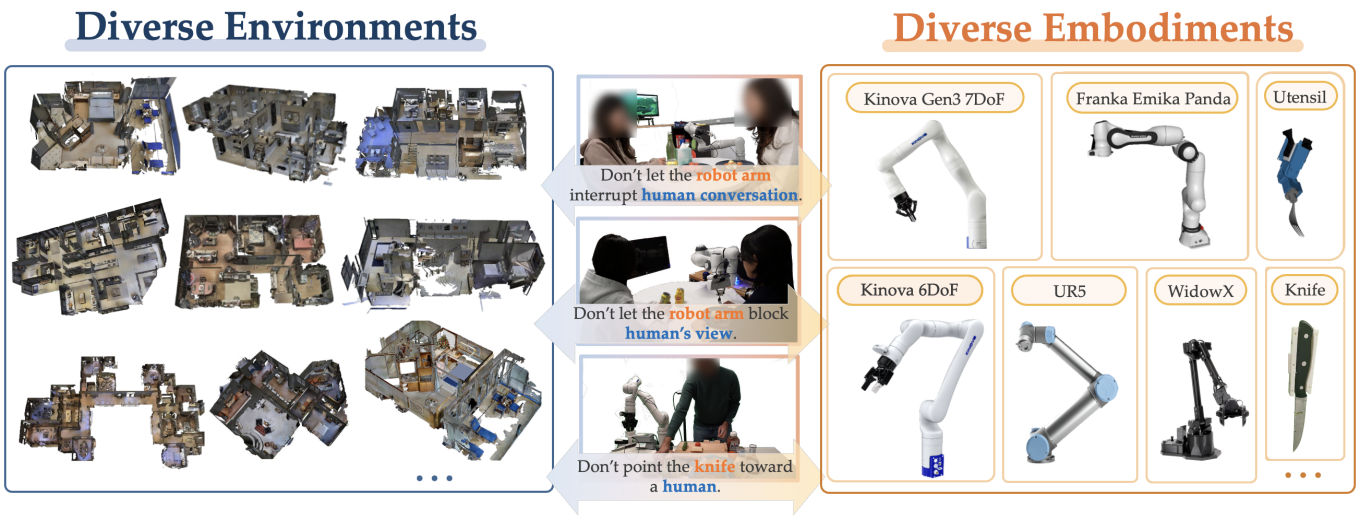


Fig. 1: Overview of E^2-CARE, which enables context-aware caregiving by modeling the joint interaction between the environment and the robot embodiment. Primitive skills are represented as reusable interaction templates and reshaped online via task-specific constraints, allowing safe deployment across diverse environments and robot embodiments.

***Abstract*—Physical caregiving robots need to assist different users with different tasks in diverse environments, and they come in many embodiments. While substantial progress has been made on individual caregiving tasks, most existing systems remain tightly coupled to specific environments and robot embodiments, and often do not explicitly model or constrain interactions around people, despite humans being special agents in the environment. This motivates a focus on adapting to context that emerges from the joint interaction between the environment and the robot's embodiment. We propose E^2-CARE, a framework that enables context-aware adaptation by representing primitive caregiving skills as interaction templates whose execution is reshaped online. E^2-CARE represents the environment, the robot, and the human within a unified 3D dynamic scene graph that models these interaction contexts explicitly, and synthesizes task-specific constraints to govern how each skill is executed. By enforcing these constraints at runtime, the same skill templates can be reused zero-shot and safely across diverse environments and robot embodiments. We evaluate E^2-CARE across four activities of daily living in hundreds of simulated household environments, including assistive home settings, and across diverse robot embodiments, and validate it through user studies on two caregiving tasks with two robots in various real-world environments. Results demonstrate consistent and successful adaptation across these environments and embodiments. Website: https://emprise.cs.cornell.edu/e2care**

I. Introduction

Consider a robot assisting people with limited mobility throughout the day: preparing breakfast in a cluttered kitchen, feeding the user at a dining table, and later assisting with grooming beside a wheelchair in a bathroom. Across these activities, robots must help users with different tasks in diverse environments. They may also come in many embodiments, and even change embodiments as they switch tools. In these settings, adaptation is not only about handling new environments or robot embodiments; it is about how the robot's embodiment interacts with the environment in the human's presence. While substantial progress has been made on individual physical caregiving tasks, most existing systems remain tightly coupled to specific environments or robot embodiments and require extensive manual reconfiguration as contexts change [33, 34, 18, 19]. Moreover, many systems do not explicitly model interactions involving people, even though humans are special agents in the environment. Crucially, safe and effective caregiving cannot be achieved by environment understanding or embodiment reasoning alone. For example, when a robot holds a knife near a person, neither "a human is nearby" nor "the robot holds a sharp tool" is sufficient: risk depends on their joint configuration. A sharp end-effector oriented toward a person is dangerous, while the same tool oriented away may be safe. This highlights the need to model context arising from joint interaction between the environment and robot embodiment in the presence of humans.

Existing works have demonstrated the ability to perform a

wide range of manipulation tasks and to adapt across environments and robot embodiments, using approaches such as task-and-motion planning [26, 25], modular skill libraries [48, 44, 24], and more recently, large-scale learning-based models [58, 22]. Notable examples include Vision-Language-Action models [17], which have shown promising performance on tasks such as cleaning kitchens and folding laundry. However, despite these successes, deploying VLA models in real-world, human-centered settings remains challenging. Most existing training datasets are collected in environments without human presence, leaving it unclear how such models generalize to settings that are less structured, involve more dynamic interactions, and impose more stringent safety requirements.

A fundamental limitation of data-driven approaches to physical caregiving is that *the individuals who most require robotic assistance are the least represented in existing datasets*. Collecting large, diverse robot datasets is expensive, time-consuming, and difficult to scale through physical deployment. For example, RT-1 [4] required approximately 130,000 demonstrations collected over 17 months, and achieving dataset sizes comparable to foundation models in other domains (e.g., large language models) would require millions of robot-hours [21, 27, 12, 37]. This challenge is further amplified in physical caregiving scenarios, which inherently require large-scale human-robot interaction data. Caregiving tasks such as feeding and dressing involve continuous interaction with humans, making data collection risky and difficult to scale. While a small number of datasets capture how human caregivers perform ADLs [29], none capture the full complexity of physical human-robot interaction. As a result, developing foundation models for physical caregiving faces challenges.

Our key insight is that robust caregiving behavior cannot be specified by the environment or the robot embodiment alone, but must be grounded in context that arises from how a particular embodiment interacts with a particular environment in the presence of humans. As a result, caregiving skills should not be treated as fixed policies, but as adaptable interaction templates whose execution depends on the current environment-embodiment context. Building on this insight, we propose E^2-CARE, a framework that explicitly represents the environment, the robot embodiment, and the human within a unified 3D dynamic scene graph. Given this representation of contexts, E^2-CARE synthesizes constraints that shape how each skill is carried out at runtime. For example: do not point the knife blade toward a human. We distinguish between hard constraints, which encode safety-critical conditions that must always be satisfied to guarantee human safety, and soft constraints, which capture task-specific requirements and are satisfied whenever possible. We enforce hard constraints and maximize satisfaction of soft constraints during execution using control barrier functions, enabling the same skill templates to be safely reused across diverse environments and robot embodiments without retraining.

We evaluate E^2-CARE across four activities of daily living (ADLs) in hundreds of simulated household environments with diverse robot embodiments, and further validate it through user studies on two real-world caregiving tasks using two robot embodiments. Our results demonstrate that explicit reasoning over environment-embodiment interaction is critical for safe and effective adaptation in physical caregiving scenarios.

II. Related Work

Environmental Adaptation of Physical Caregiving Robots. Prior work has developed physical caregiving robots for a variety of activities of daily living (ADLs), including feeding [18, 36, 13], grooming [56, 9], dressing [45, 54, 8, 57], bathing [59, 34, 16, 30], and meal preparation [10][52]. These systems typically achieve reliable performance by designing task-specific pipelines that are often tightly coupled to a particular environment [19, 34, 53]. These assumptions fundamentally limit generalization. When deployed in real-world household environments with different room geometries, furniture types, object arrangements, or human presence, these systems often require extensive manual reconfiguration [36, 19, 18]. As a result, existing physical caregiving robots struggle to adapt zero-shot to the physical and social characteristics of unseen household environments. In contrast, we explicitly reason about environmental structure where humans are treated as special agents, allowing caregiving skills to adapt at execution time rather than relying on environment-specific pipelines.

Embodiment Adaptation of Physical Caregiving Robots. Existing work has demonstrated physical caregiving across multiple ADLs on a variety of robot embodiments [40, 14, 39, 1, 2, 49, 38]. However, most systems are designed around a single robot embodiment and do not explicitly model how embodiment differences affect task execution [18]. Consequently, when a system is deployed on a different robot, manual modification is often required. In these systems, tools used for caregiving are typically treated as interchangeable end effectors rather than as components that fundamentally alter the robot's effective embodiment and its interaction constraints (e.g., sharp tools near humans or a wet sponge near electronics). As a result, embodiment-dependent execution constraints are rarely modeled explicitly and are instead encoded implicitly through task-specific design choices. Beyond technical system design, prior work has shown that robot embodiment strongly influences user acceptance in caregiving contexts, and that assistance needs evolve over time as users age [47, 46, 20]. Together, these findings highlight that embodiment is central to both physical interaction and long-term caregiving deployment. Our work explicitly models robot embodiment and its interaction with the environment and humans, enabling skills to be reused across embodiments.

Adaptation to Environment and Embodiment in Robot Manipulation. Prior work addresses both environmental and embodiment variation by training generalist manipulation policies on large, diverse datasets spanning multiple scenes and robot embodiments [3, 22, 17, 58]. While promising, these data-driven methods typically rely on large-scale supervision and require substantial amounts of data and computation [28, 7, 50, 11]. Moreover, environment and embodiment are often

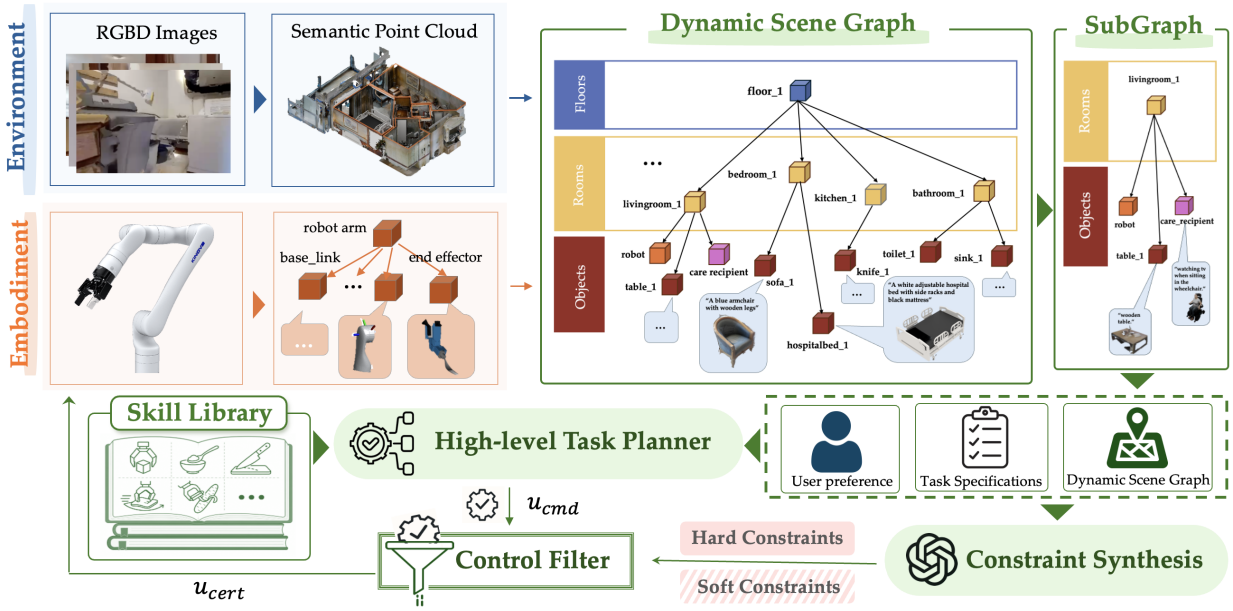


Fig. 2: E^2-CARE framework. We represent the environment, the robot, and the human within a unified 3D dynamic scene graph that models these interaction contexts explicitly, and synthesizes task-specific constraints to govern how each skill is executed. By enforcing these constraints at runtime, the same skill templates can be reused zero-shot and safely across diverse environments and robot embodiments.

handled implicitly via data coverage, rather than being explicitly modeled. Other work explicitly represents aspects of the environment to support task execution [5, 42]. However, these approaches typically focus on static environments and consider only a limited set of objects, either ignoring human presence or predefining fixed safety distances from humans. Consequently, they remain insufficient for physical caregiving scenarios, where robots must continuously operate in close proximity to humans. Our work addresses this gap by explicitly reasoning about environment-embodiment interactions in the presence of a human, conditioned on the task, enabling existing caregiving skills to be modulated at execution time without policy retraining or dataset expansion.

III. Problem Formulation

We consider the problem of **embodiment- and environment-aware physical caregiving robots** performing *activities of daily living* (ADLs) in diverse human living environments. The robot operates with an embodiment $\mathcal{R}$. The embodiment may change over time as the robot equips or removes tools (e.g., a utensil for feeding or a knife for meal preparation). We assume $\mathcal{R}$ is specified by a Unified Robot Description Format (URDF), defining the robot's kinematics, degrees of freedom, and tool attachments. The robot executes tasks within an environment $\mathcal{E}$ that varies in spatial layout, furniture geometry, object configurations, and human presence. We assume access to the primitive skill library $\mathcal{L} = \{l_1, l_2, ...l_k\}$, consisting of parameterized skills (e.g., *navigate_to*(p) for $p \in \mathbb{R}^3$) and learned skills (e.g., *cutting*), which can be composed to perform ADL tasks. The central challenge is to execute caregiving skills for activities of daily living safely and effectively across variations in both $\mathcal{R}$ and $\mathcal{E}$. Given an ADL task τ and a user preference expressed in natural language, the robot receives an observation o_t at timestep t, where $o_t = (i_t, f_t, p_t)$, with $i_t \in \mathbb{R}^{W \times H \times 4}$ an RGB-D image, $f_t \in \mathbb{R}^6$ the force-torque readings ($\mathbf{0}_6$ if unavailable), and p_t the joint state of the robot. At timestep t, the robot selects a primitive skill $l_t \in L$ and executes an action $a_t \in l_t$.

IV. E^2-CARE

We present E^2-CARE, a framework for environment-embodiment-aware physical caregiving robots (Figure 2). E^2-CARE models the environment, the robot, and the human within a unified 3D dynamic scene graph to explicitly represent their geometric and semantic interactions from sensory observations. Given a task τ and a skill library $\mathcal{L}$, a Planning Domain Definition Language (PDDL)-based planner generates a sequence of primitive skills using symbolic predicates grounded in the scene graph. Conditioned on the task τ and user preference, an LLM-based reasoning module performs semantic reasoning to analyze the scene graph and derive embodiment- and environment-dependent interaction constraints for each skill, including hard safety constraints and soft task preferences. These constraints are converted into state-dependent control conditions, where hard constraints are enforced at execution time through a quadratic-program (QP) control filter based on control barrier functions (CBFs), while soft constraints guide optimization when feasible. This design enables the same skill primitives to be reused safely and efficiently across diverse environments and robot embodiments by reshaping their execution through context-dependent constraints. We detail each component below.

A. 3D Dynamic Scene Graph Generation

3D Dynamic Scene Graph $G(V_t, E_t)$ at time step t is a layered, hierarchical representation that abstracts dense 3D reconstructions into higher-level spatial concepts such as objects, agents (including the robot and humans), and rooms, while explicitly modeling their spatio-temporal relationships (e.g., "object A is in room B at time t"). This structure enables the system to reason over embodiment-environment interactions using semantic entities rather than raw geometry alone.

Each node $i \in V_t$ corresponds to a spatial entity (e.g., a manipulable object, a semantic region, the robot, or a human) and stores attributes (p_i, f_i), where p_i is a geometric representation (e.g., a point cloud) and f_i is a semantic descriptor. Edges encode pairwise spatial and semantic relations such as adjacency, containment, and contact. Nodes are updated over time through multi-frame association and tracking, allowing the graph to maintain persistent identities and support dynamic reasoning as the scene evolves.

a) Environment: We construct the environment graph using a SLAM pipeline [32, 43, 15]. At each discrete time step, the robot receives an observation (i_t, u_t), where $i_t \in \mathbb{R}^{W \times H \times 4}$ is an RGB-D image and $u_t \in \mathbb{R}^6$ denotes IMU measurements, with the camera pose estimated by the SLAM backend. The backend maintains a metric map that is abstracted into semantic entities and relations. For each frame i_t, open-vocabulary object detection (RAM [55], GroundingDINO [31]) and segmentation (SAM [23]) produce a set of object masks $\{m_k^t\}$, where each mask corresponds to a detected object k. These masks are associated across frames and projected into 3D point clouds $\{p_k^t\}$ using depth and pose estimates. The resulting graph maintains a hierarchical structure: at a high level it provides a compact semantic abstraction of the environment, while at a lower level it grounds semantics into geometry, yielding a representation that supports downstream planning and execution-time reasoning.

b) Embodiment: We convert the robot's URDF into a kinematic graph that captures the structural organization of its components and add it into the 3D dynamic scene graph. Rather than treating the robot as a single rigid entity, we represent it as a set of functional components, including the mobile base, arm, and end effector, where the end effector may change as the robot equips different tools. At the geometric level, each component is associated with its mesh model and a collision proxy representation. Following [35], we approximate each link using a set of collision spheres anchored to the kinematic structure, enabling efficient computation of collision checking during execution. The robot embodiment is represented hierarchically: at a high level, it provides a functional abstraction of the robot (mobile base, arm, end effector); at a lower level, it encodes link-level geometry, yielding a grounded representation that supports robot behavior adaptation.

B. High-Level Planning

Given a task τ, a skill library L (see Appendix), and the current 3D dynamic scene graph $G(V_t, E_t)$, we first query an LLM to identify task-relevant entities and prune irrelevant nodes, producing a task-conditioned subgraph $G(V_t', E_t')$. For example, during a feeding task, nodes corresponding to unrelated regions (e.g., a bathroom) are removed. This subgraph preserves the semantic and geometric structure necessary for the task and serves as the environment-embodiment context while reducing reasoning complexity. We then ground the task-relevant subgraph $G(V_t', E_t')$ into a symbolic PDDL state by mapping nodes to objects, and perform high-level planning using a PDDL-based planner to generate a task plan that achieves the goal.

```
# an example skill primitive
(:action Pickup
  :parameters (?obj - object)
  :precondition (and (GripperFree)
                     (Reachable ?obj))
  :effect (and (Holding ?obj)
               (not (GripperFree))
)
```

C. Constraint Synthesis

Given a PDDL-planner skill sequence, E^2-CARE treats each primitive skill as an interaction template and adapts them at execution time by synthesizing environment-embodiment interaction constraints for each skill. These constraints specify how a skill should be executed safely and appropriately in the current context, taking into account both environmental conditions and the effective embodiment of the robot.

We define two types of constraints: hard constraints and soft constraints. Hard constraints encode safety-critical states and must never be violated, such as ensuring that the robot does not collide with a human. These constraints are predefined and enforced deterministically using control barrier functions. Soft constraints encode user preferences that affect execution behavior but may be relaxed when necessary, such as avoid occluding the user's view.

a) Constraint Representation: We represent constraints using a symbolic interface that separates *what is measured* from *what rule is enforced*. For example, a safety constraint such as "keep the end-effector 30 cm away from a human" measures distance and enforces a minimum bound. This structure allows semantic instructions to be compiled into control-safe rules. Formally, each constraint is encoded as a tuple

$$c_i = (s_i, o_i, r_i, m_i, \theta_i), \tag{1}$$

where s_i is a robot embodiment entity (e.g., an end-effector or arm link) from $G(V_t', E_t')$ and o_i is a scene entity from $G(V_t', E_t')$ (e.g., a human or object). The relation type $r_i \in \mathcal{R}$ specifies the geometric quantity being measured, and $\mathcal{R}$ is defined as

$$\mathcal{R} = \{\text{distance}, \text{region}, \text{orientation}, \text{velocity}\}. \tag{2}$$

for all tasks. The mode m_i specifies the behavioral rule applied to that measurement, such as enforcing a lower bound, an upper bound, or directional avoidance. The parameter vector θ_i contains numeric thresholds (e.g., distance limits or angular bounds). Together, (r_i, m_i, θ_i) define both the measured quantity and the rule governing it. Every constraint therefore

corresponds to a safe set in state space that can be compiled into a differentiable control barrier function.

b) Hard Constraints: Hard constraints encode safety-critical conditions that must never be violated, such as collision avoidance, joint limits, and minimum separation between the robot and humans. These constraints are predefined to guarantee safety.

Example: Cutting with a knife

Joint limits:
$(\text{robot}, \varnothing, \text{joint}, \text{inside-limit}, \text{joint\_limit\_bounds})$
Self-collision avoidance:
$(\text{robot link}, \{\text{other links}\}, \text{distance}, \text{keep-distance}, 0.05)$
Human safety distance:
$(\text{robot}, \{\text{human}\}, \text{distance}, \text{keep-distance}, 0.1)$

In general, we define x as the current state of the robot, the human, and objects from the dynamic scene graph. A constraint is represented as a control barrier function $h(x)$: $h(x) \geq 0$ indicates safe states, while $h(x) < 0$ corresponds to unsafe states. For example, a human-robot distance constraint can be written as $h(x) = \text{dist}(\text{robot}, \text{human}) - d_{\min}$ where $d_{\min}$ is the minimum distance threshold. Formally, each hard constraint is represented as a control barrier function $h_{\text{hard}}(x) \geq 0$.

c) Soft Constraints: Soft constraints encode semantic, social, or user preferences that should be satisfied when feasible but may be relaxed to preserve safety or feasibility. For example, the robot should avoid blocking a user's view while they are watching TV.

Example: Feeding while the user is watching TV

$(\text{robot}, \{\text{human, tv}\}, \text{region}, \text{outside-region}, 0.1)$
$(\text{robot}, \varnothing, \text{velocity}, \text{limit-speed}, 0.1)$

Soft constraints are generated by querying an LLM-based reasoner with the task specification, user preferences, and the task-relevant subgraph $G(V_t', E_t')$. The reasoner outputs symbolic constraint tuples, which are grounded using geometric attributes stored in the scene graph. Each grounded constraint is compiled into a differentiable barrier function

$$h_{\text{soft}_i}(x) \geq 0, \tag{3}$$

where violation corresponds to $h_{\text{soft}_i}(x) < 0$.

D. Constraint-Aware Skill Execution

For each skill, given an associated set of hard and soft constraints, execution is regulated using an operational-space control barrier function (OSCBF) filter [35]. The filter strictly enforces hard constraints while biasing execution toward satisfying soft constraints whenever feasible. This allows the robot to adapt skill execution to the current context while preserving task intent and guaranteeing safety. We model the robot with control-affine dynamics in joint space

$$\dot{q} = u, \tag{4}$$

where q denotes joint configuration and u is the commanded joint velocity. Constraints are defined in task space using geometric quantities extracted from the scene graph. Let

$$x = FK(q) \tag{5}$$

denote the task-space state (e.g., end-effector pose, relative distance to a human, or orientation), obtained through forward kinematics. Each constraint is represented as a control barrier function $h(x) \geq 0$, where $h(x) < 0$ corresponds to unsafe or undesirable states. Using the chain rule,

$$\dot{h}(x, u) = \nabla h(x)\, J(q)\, u, \tag{6}$$

where $J(q)$ is the task Jacobian mapping joint velocities to task-space velocities. Enforcing

$$\dot{h}(x, u) \geq -\alpha(h(x)) \tag{7}$$

guarantees forward invariance of the safe set, where α is an extended class $\mathcal{K}_\infty$ function.

The certified control u_{cert} is obtained by solving a quadratic program. The objective encourages tracking of the nominal command u_{cmd} produced by the primitive skill. Slack variables s_i allow soft constraints to be relaxed when necessary while penalizing violations:

$$\begin{aligned} u_{\text{cert}} = \arg \min_{u \in \mathcal{U},\, s_i \geq 0} \quad & \|u - u_{\text{cmd}}\|_2^2 + \sum_i s_i^2 \\ \text{s.t.} \quad & \nabla h_i^{\text{soft}} J u \geq -\alpha(h_i^{\text{soft}}) - s_i, \quad \forall i \\ & \nabla h_j^{\text{hard}} J u \geq -\alpha(h_j^{\text{hard}}), \qquad \forall j \\ & \|u - u_{\text{cmd}}\|_2 \leq \delta_u. \end{aligned} \tag{8}$$

Here δ_u defines the maximum allowable deviation from the nominal control command. The resulting controller acts as a safety filter that modifies each primitive skill only as much as necessary to satisfy environment-embodiment constraints. This guarantees certified safety while preserving the structure and intent of the planned skill sequence, enabling context-aware execution without retraining or task-specific controller design.

V. Simulation Experiments

In this section, we address the following research questions through extensive simulation experiments, which provide a safe and scalable evaluation environment:

Q1. Can our framework identify environment-embodiment interaction constraints for caregiving tasks?

Q2. Can our system adapt caregiving primitive skills across different environments and robot embodiments based on the environment-embodiment interaction context?

A. Simulation Experimental Setup

We evaluate our framework across diverse robot embodiments, household environments, and caregiving tasks. We consider five different robot embodiments: a Kinova Gen3 (7-DoF), a Kinova 6-DoF arm, a Franka Emika Panda, a UR5, and a WidowX 250. For environments, we evaluate in 130 simulated household [6, 41] settings with varying furniture layouts and object types, including 30 assistive homes [51]. We assess performance across multiple activities of daily living (ADLs), including feeding, bathing, grooming, and meal preparation. For each scenario, ground-truth constraint annotations are labeled by three third-party human annotators.

Baselines. We evaluate both internal ablations and external baselines. We include ablations: (i) **Ours w/ GT perception**, where perception modules are replaced with ground-truth state, establishing an upper bound on performance under perfect scene understanding; (ii) **Ours w/o dynamic scene graph**, which removes the unified graph representation, representing embodiment and environments with natural language; (iii) **Ours w/o constraints**, where fixed primitive skill templates are executed without adapting to constraints.

We compare against methods that are closely related to our work and explicitly model the environment to ensure safety. Since learning-based models for physical caregiving robots are not yet sufficiently mature due to limited data availability (Section I), we focus our comparisons on model-based, safety-constrained state-of-the-art methods: (i) **SemanticSafe** [5], which represents the environment as an unstructured object list and encodes all the constraints as barrier functions; (ii) **Nominal CBF** [35], which enforces a fixed set of pre-specified safety constraints without modeling environment structure.

Metrics. We evaluate two primary metrics, following prior work [35, 5]: (1) *Task success rate*, defined as the percentage of episodes in which the caregiving task is completed without collision; and (2) *Constraint satisfaction rate*, defined as the fraction of timesteps in which all constraints are satisfied.

B. Evaluating Constraint Synthesis

We first evaluate whether our framework can correctly identify environment-embodiment interaction constraints given the context. To isolate this component, we provide ground-truth dynamic scene graphs and evaluate the synthesized constraints. Table I reports the precision and recall of valid constraint synthesis across four ADL tasks. Our system achieves consistently high recall (around 90%) and strong precision across tasks, indicating that the framework reliably identifies environment-embodiment constraints. For example, in a meal-preparation task, the robot imposes orientation constraints on the knife's pose during cutting when a human is nearby. Tasks involving more complex spatial interactions, such as grooming, exhibit slightly lower precision. Overall, these results demonstrate that the unified environment-embodiment representation enables reliable constraint generation across diverse contexts.

TABLE I: Constraint Synthesis Precision and Recall on 4 ADLs across Household Environments.

Task	Precision	Recall
Feeding	91.0% ± 5.1%	94.8% ± 5.1%
Meal Prep.	88.2% ± 10.9%	92.1% ± 9.5%
Grooming	86.0% ± 8.7%	89.0% ± 9.5%
Bathing	90.3% ± 5.8%	90.0% ± 10.9%

C. Evaluating Adaptation to Environments

To study how well our system adapts to diverse environments, we evaluate it across 130 household environments, including 30 assistive homes equipped with assistive devices such as hospital beds and Hoyer lifts, which introduce additional environmental complexity. Table II and Figure 3 report

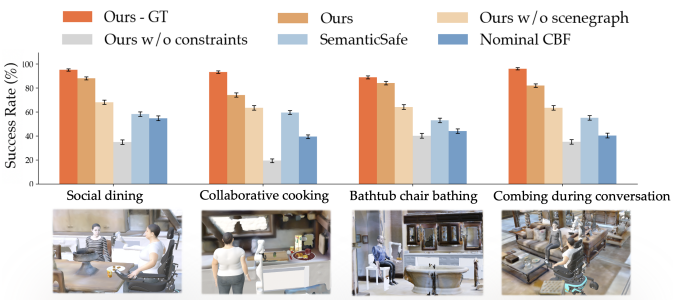


Fig. 3: Success rate across different scenarios (Table II) on the Kinova Gen3 robot in simulation (More in Appendix).

the success rate and average constraint satisfaction rate across eight scenarios spanning feeding, meal preparation, bathing, and grooming tasks with various human activities, e.g., solitary feeding or social dining (scenario details are shown in Table II). Our method consistently achieves the highest success rate (around 95%) and the highest constraint satisfaction rate (80%) across all scenarios, closely approaching the upper bound achieved by Ours w/ GT perception, demonstrating our robustness to imperfect perception.

Our method maintains high reliability, achieving between 82.8% and 96.1% constraint satisfaction rate, with an average gap of less than 6% from the ground-truth perception upper bound. Removing the shared scene graph leads to a substantial degradation in performance. Ours w/o dynamic scene graph drops by about 15 percentage points in success rate across most scenarios. Without explicitly modeling environment-embodiment interaction, the system frequently ignores how the held tool affects the surrounding environment. For example, the robot may point a knife blade toward a human in a social dining context (C2) or move a wet sponge above an electrical device during bathing (C5). Ours w/o constraints consistently performs the worst among ablations, often falling below 45% in complex interaction scenarios. It ignores human presence in the environment and often causes the robot arm to move dangerously close to, or even collide with, a person. This confirms that fixed skill templates alone are insufficient for safe adaptation to diverse environments.

Compared to the baselines, our method shows a clear advantage. SemanticSafe and Nominal CBF achieve moderate constraint satisfaction rate but remain 15–40 percentage points below our full system in most scenarios. These methods lack a structured representation of the environment. As a result, SemanticSafe often hallucinates constraints between objects and ignores human presence. For example, it may generate a constraint to regulate knife orientation but fail to account for human position, causing the blade to point toward a person. Without explicit modeling of interaction context, the system often hallucinates multiple constraints that make the problem infeasible under the resulting control barriers and frequently stops at a hallucinated barrier.

Overall, the simulation results show that our framework can reliably synthesize environment-embodiment interaction constraints and adapt caregiving skills across environmental

TABLE II: Constraint satisfaction rate of the Kinova Gen3 robot performing four ADLs across diverse scenarios.

Task	Context (C.)	Ours w/ GT Perception	Ours	Ours w/o graph	Ours w/o constraints	SemanticsSafe [5]	Nominal CBF [35]
Feeding	C1. Dining while the care recipient is watching television	100.0% ± 0.0%	**96.1% ± 6.3%**	68.3% ± 7.5%	34.8% ± 14.2%	54.3% ± 12.7%	42.3% ± 23.1%
	C2. Social dining with others	95.2% ± 9.7%	**90.0% ± 14.1%**	54.3% ± 19.7%	35.8% ± 28.3%	46.1% ± 25.9%	42.1% ± 28.9%
Meal Preparation	C3. Cooking near the care recipient	100.0% ± 0.0%	**92.0% ± 3.4%**	77.0% ± 18.4%	45.8% ± 24.4%	72.0% ± 26.3%	69.2% ± 29.5%
	C4. Cooking with a caregiver present	89.1% ± 8.4%	**82.8% ± 12.7%**	67.1% ± 17.8%	30.7% ± 10.5%	59.4% ± 12.8%	45.8% ± 23.4%
Bathing	C5. Bathtub chair bathing	83.1% ± 7.8%	**75.2% ± 11.9%**	56.3% ± 27.2%	33.6% ± 24.2%	55.8% ± 12.4%	54.6% ± 13.7%
	C6. Bed bathing	73.5% ± 13.2%	**65.9% ± 12.7%**	45.2% ± 19.9%	41.6% ± 17.3%	42.5% ± 18.1%	43.7% ± 16.9%
Grooming	C7. Combing while the care recipient is watching television	96.6% ± 5.3%	**87.0% ± 7.7%**	64.1% ± 19.6%	38.4% ± 16.4%	65.4% ± 14.6%	59.8% ± 13.4%
	C8. Combing while the care recipient is talking to the caregiver	95.7% ± 4.6%	**85.3% ± 8.9%**	45.7% ± 29.1%	33.3% ± 18.1%	55.3% ± 26.9%	47.3% ± 25.0%

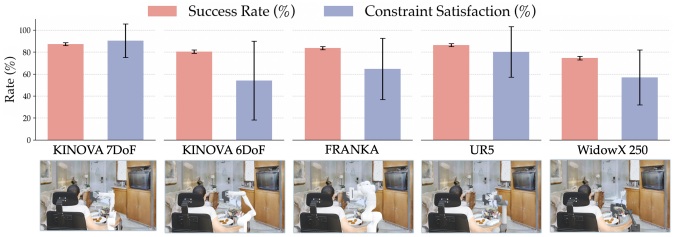


Fig. 4: Success rate and constraint satisfaction rate of ADL tasks on different robots.

variation, achieving both high task success and a high constraint satisfaction rate.

D. Evaluating Adaptation to Embodiments

Figure 4 evaluates our method on different robot embodiments. Our framework maintains strong performance across all platforms, demonstrating robustness to embodiment variations. The Kinova Gen3 achieves the highest performance, with approximately 87% success rate and 90% constraint satisfaction rate. The UR5 achieves comparable results (86% success rate and 80% constraint satisfaction rate), suggesting that the synthesized constraints generalize well across industrial manipulators with different link geometries. Overall, performance remains high across embodiments despite substantial differences in morphology.

Even on more constrained embodiments such as the WidowX250 and Kinova 6-DoF, which have more restrictive kinematic limits, the system maintains about 75% success and 55% constraint satisfaction rate. The performance drop is expected due to embodiment limitations: the controller continues to enforce constraints whenever feasible, but certain constraints become physically difficult to satisfy. For example, the Franka robot's larger physical footprint makes it harder to avoid blocking the user's view during feeding. Importantly, the relatively small performance gap between embodiments highlights a key property of our approach: skills represented as interaction templates adapt through embodiment-environment-aware constraint synthesis. These results support our central claim that explicitly modeling embodiment-environment interaction enables adaptive caregiving behaviors.

VI. Real-world User Study

In this section, we further evaluate our approach on real robotic systems via a Cornell IRB-approved real-world user study measuring perceived safety, behavioral appropriateness, and user satisfaction. Specifically, we address the following research question: Do users perceive that the robot adapts its behavior to the environmental context, and do they feel safe during task execution?

A. Real-World Experimental Setup

Hardware. We evaluate our framework on two robot embodiments: (i) a Kinova Gen3 7-DoF robotic arm and (ii) a Franka Emika Panda 7-DoF robotic arm. Each robot is equipped with an Intel RealSense D435i RGB-D camera mounted on the wrist for egocentric visual perception. In addition, we use an external Intel RealSense D435i camera to provide a third-person global view. We incorporate force feedback similar to prior work on physical caregiving systems [19]. In contact-rich scenarios (e.g., bite transfer during feeding), we switch to a task-space compliant controller that uses force feedback to regulate interaction. We also enforce force thresholds for safety; when measured forces exceed predefined limits, the system triggers an emergency stop.

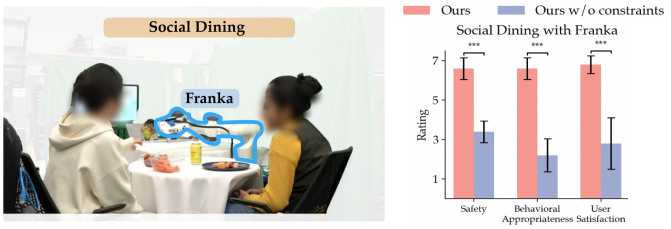
Evaluation Scenarios. We evaluate our approach on two real-world activities of daily living (ADLs): feeding and meal preparation. For feeding, we use a parameterized skill library in [19]. For meal preparation, we collected 300 demonstrations of cutting and peeling and fine-tuned Pi0.5 to obtain Vision-Language-Action (VLA)-based primitive skills for these behaviors. Our framework supports both learned and parameterized skills and is agnostic to how skills are obtained. We conduct a user study with five participants per scenario, evaluating both tasks across varied environments and two robot embodiments. For safety reasons, we exclude the no-constraint baseline from the meal-preparation experiments involving sharp tools.

Baselines. We compare two methods in the user study: (i) our full system and (ii) our system without constraints synthesis and satisfaction.

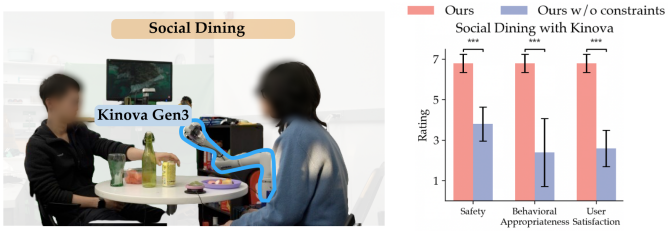
User Study. We recruit five participants per scenario (3 female, 2 male; ages 20-26). Participants completed consent and demographic forms and received standardized instructions prior to the study. Participants then physically interacted with the robot in each condition in a counterbalanced order and completed a post-task evaluation. Participants evaluated each interaction using a 7-point Likert scale across 3 criteria: perceived safety, behavioral appropriateness, and user satisfaction.

B. Results

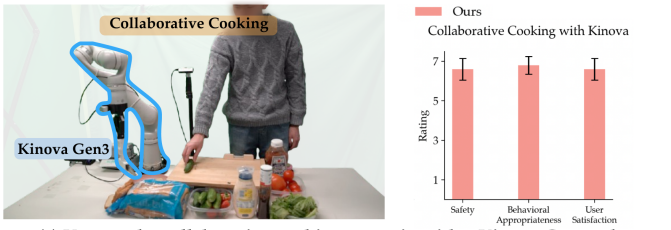
Results are shown in Figure 5. In social dining scenarios with both the Kinova and Franka robots, our method significantly outperforms the baseline across all three metrics. The



(a) User study: social dining scenario with a Kinova Gen3 robot



(b) User study: social dining scenario with a Franka robot



(c) User study: collaborative cooking scenario with a Kinova Gen3 robot

Fig. 5: User study across diverse environments with Franka and kinova robots in social dining and collaborative cooking.

largest improvement appears in behavioral appropriateness, suggesting that our framework successfully enables the robot to adapt its behavior to the current context and demonstrating robustness across embodiments. In the collaborative cooking scenario, our method maintains high ratings across all dimensions. The robot explicitly adapts its behavior based on environment-embodiment interaction. For example, when the robot is holding a sharp knife and a human is approaching, it rotates the blade away from the human and maintains a safe distance (shown in Fig. 6); in contrast, when the robot is equipped with a standard gripper or the human is far away, no specific orientation constraint is imposed. Participants consistently perceived the robot as safe and appropriate in context, even when the robot was holding a sharp knife. One participant remarked: "What made the robot feel safe was not just that it avoided me, but that it seemed aware of my movement and adjusted the knife orientation in response."

C. Runtime analysis

We report runtime across 20 trials on a workstation (RTX 4070 GPU; Intel i7-13700 CPU; 32GB RAM): *(i) Perception*: Scene graph construction via open-world detection takes 1.15 ± 0.39s (once per skill). A lightweight human detector runs in the control loop at $(3.85 \pm 0.14) \times 10^{-3}$s for rapid response to human motion. *(ii) Planning*: The LLM prunes irrelevant nodes in 1.94 ± 0.36s, and then the PDDL planner generates the skill sequence in $(1.5 \pm 0.2) \times 10^{-4}$s. Constraint synthesis takes 1.96 ± 0.52s and is performed only once per skill (not per timestep). *(iii) Control*: The CBF-based QP filter has negligible overhead $(1.5 \pm 0.3) \times 10^{-4}$s, remaining under 3×10^{-4}s even with 1000 constraints, ensuring real-time safety enforcement. When a human suddenly moves, our method detects the motion in about 0.0004s and reacts within 0.0002s during execution.

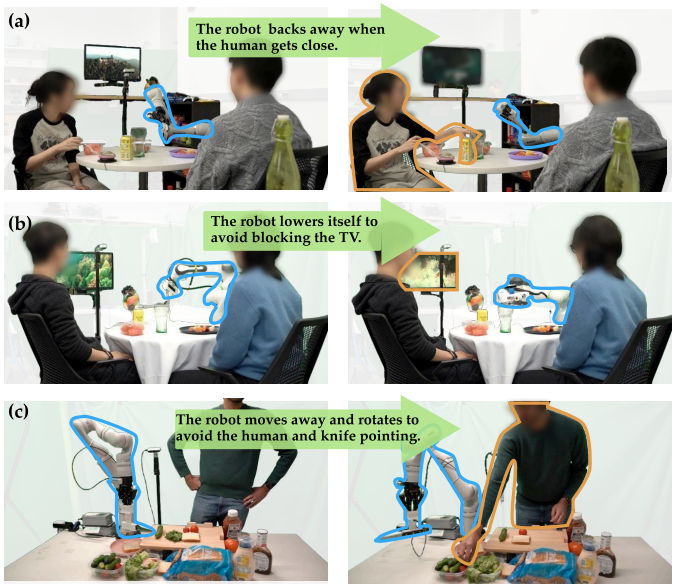


Fig. 6: Qualitative results. In (a), while the robot is feeding a human in a social dining scenario, another person approaches the table to get food, and the robot backs away to maintain a safe distance. In (b), the robot lowers its motion to avoid blocking the user's view and then moves away to maintain a safe distance when both attempt to reach the same food. In (c), when a human approaches a robot holding a knife, the robot immediately increases the separation distance and rotates the knife blade away for safety. Once the human leaves the workspace, the robot resumes the task.

VII. Discussions and Limitations

Our results show that context-aware caregiving behavior emerges from the joint interaction between the environment and the robot embodiment, rather than from either in isolation. In human-centered settings, executing a nominally correct skill without accounting for this interaction can lead to unsafe or socially inappropriate behavior when humans, robots, and surrounding objects co-exist. By explicitly synthesizing environment-embodiment interaction constraints grounded in a shared scene graph, we enable the same skill to be executed differently depending on context, resulting in higher constraint satisfaction in simulation and improved perceived safety and behavioral appropriateness in user studies. Our framework supports both learned and parameterized skills and is agnostic to how skills are obtained. We include a VLA policy to demonstrate compatibility with learned skills, while parameterized skills from existing systems can be integrated without additional data collection. Once a skill is available, it can be reused zero-shot across diverse environments and robot

embodiments.

Despite these benefits, the framework has several limitations. First, constraint synthesis relies on prior knowledge encoded in vision-language models (VLMs), and the quality of the generated constraints is therefore tied to the VLM's understanding of the context; misinterpretations or omissions may lead to missing or overly conservative constraints. Second, our framework depends on the quality of perception and scene representation. We provide controller-level safety, but end-to-end safety is conditional on perception accuracy and propagated module errors. Errors in object detection, semantic labeling, or dynamic scene graph construction can propagate to constraint synthesis and affect execution. Finally, the scope of adaptation is limited by the predefined primitive skill library. While treating skills as adaptable interaction templates enables reuse across environments and robot embodiments, behaviors outside the coverage of available skills cannot be executed. Supporting online skill discovery or learning new primitives would further improve flexibility and task coverage.

Our framework currently focuses on safety constraints that must always hold throughout skill execution. One promising direction for future work is to extend the framework to support more expressive temporal constraints. For example, incorporating richer temporal specifications from formal methods such as Signal Temporal Logic (STL), including operators such as *Eventually* and *Until*, would enable reasoning about more complex temporal behaviors and further improve the expressiveness of the framework.

Acknowledgments

This work was partly funded by National Science Foundation IIS #2132846, and CAREER #2238792. This research was also funded, in part, by the Advanced Research Projects Agency for Health (ARPA-H) Agreement No. 140D042590012. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

[1] Meet Obi. https://meetobi.com/. [Online; accessed 1-Jan-2026].

[2] Neater eater robot, 2024. URL https://www.neater.co.uk/neater-eater-robotic. (Accessed: 1st January, 2026).

[3] Kevin Black et al. π_0: A vision-language-action flow model for general robot control, 2025. URL https://arxiv.org/abs/2410.24164.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*, 2023.

[5] Lukas Brunke, Yanni Zhang, Ralf Römer, Jack Naimer, Nikola Staykov, Siqi Zhou, and Angela P. Schoellig. Semantically safe robot manipulation: From semantic scene understanding to motion safeguards. *IEEE Robotics and Automation Letters*, 10(5):4810–4817, 2025.

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[7] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *Robotics: Science and Systems XIX*, 2024.

[8] Alexander Clegg, Zackory Erickson, Patrick Grady, Greg Turk, Charles C. Kemp, and C. Karen Liu. Learning to collaborate from simulation for robot-assisted dressing. *IEEE Robotics and Automation Letters*, 5(2):2746–2753, 2020. doi: 10.1109/LRA.2020.2972852.

[9] Nathaniel Dennler, Eura Shin, Maja Mataric, and Stefanos Nikolaidis. Design and evaluation of a hair combing system using a general-purpose robotic arm. pages 3739–3746, 09 2021. doi: 10.1109/IROS51168.2021.9636768.

[10] Chenyu Dong, Liangliang Yu, Masaru Takizawa, Shunsuke Kudoh, and Takashi Suehiro. Food peeling method for dual-arm cooking robot. In *2021 IEEE/SICE International Symposium on System Integration (SII)*, pages 801–806, 2021. doi: 10.1109/IEEECONF49454.2021.9382700.

[11] Cunxin Fan, Xiaosong Jia, Yihang Sun, Yixiao Wang, Jianglan Wei, Ziyang Gong, Xiangyu Zhao, Masayoshi Tomizuka, Xue Yang, Junchi Yan, et al. Interleave-vla: Enhancing robot manipulation with image-text interleaved instructions. In *The Fourteenth International Conference on Learning Representations*, 2025.

[12] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[13] Daniel Gallenberger, Tapomayukh Bhattacharjee, Youngsun Kim, and Siddhartha S Srinivasa. Transfer depends on acquisition: Analyzing manipulation strategies for robotic feeding. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 267–276. IEEE, 2019.

[14] Jennifer Grannen, Yilin Wu, Suneel Belkhale, and Dorsa Sadigh. Learning bimanual scooping policies for food acquisition. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=qDtbMK67PJG.

[15] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and*

Automation (ICRA), pages 5021–5028. IEEE, 2024.

[16] Yijun Gu and Yiannis Demiris. Vttb: A visuo-tactile learning approach for robot-assisted bed bathing. *IEEE Robotics and Automation Letters*, 9(6):5751–5758, 2024.

[17] Physical Intelligence. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

[18] Rajat Kumar Jenamani, Priya Sundaresan, Maram Sakr, Tapomayukh Bhattacharjee, and Dorsa Sadigh. Flair: Feeding via long-horizon acquisition of realistic dishes. *arXiv preprint arXiv:2407.07561*, 2024.

[19] Rajat Kumar Jenamani, Tom Silver, Ben Dodson, Shiqin Tong, Anthony Song, Yuting Yang, Ziang Liu, Benjamin Howe, Aimee Whitneck, and Tapomayukh Bhattacharjee. Feast: A flexible mealtime-assistance system towards in-the-wild personalization. In *Robotics: Science and Systems (RSS)*, 2025.

[20] Charles C Kemp, Aaron Edsinger, Henry M Clever, and Blaine Matulevich. The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3150–3157. IEEE, 2022.

[21] Alexander Khazatsky et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[22] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[24] George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'09, page 1015–1023, Red Hook, NY, USA, 2009. Curran Associates Inc. ISBN 9781615679119.

[25] Nishanth Kumar, Tom Silver, Willie McClinton, Linfeng Zhao, Stephen Proulx, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Jennifer Barry. Practice makes perfect: Planning to learn skill parameter policies. *Robotics: Science and Systems*, 2024.

[26] Nishanth Kumar, William Shen, Fabio Ramos, Dieter Fox, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Caelan Reed Garrett. Open-world task and motion planning via vision-language model generated constraints. *IEEE Robotics and Automation Letters*, 2026.

[27] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018. doi: 10.1177/0278364917710318. URL https://doi.org/10.1177/0278364917710318.

[28] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[29] Xiaoyu Liang, Ziang Liu, Kelvin Lin, Edward Gu, Ruolin Ye, Tam Nguyen, Cynthia Hsu, Zhanxin Wu, Xiaoman Yang, Christy Sum Yu Cheung, Harold Soh, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. OpenRoboCare: A Multi-Modal Multi-Task Expert Demonstration Dataset for Robot Caregiving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[30] Fukang Liu, Kavya Puthuveetil, Akhil Padmanabha, Karan Khokar, Zeynep Temel, and Zackory Erickson. Skingrip: An adaptive soft robotic manipulator with capacitive sensing for whole-limb bed bathing assistance. *arXiv preprint arXiv:2405.02772*, 2024.

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[32] Joel Loo, Zhanxin Wu, and David Hsu. Open scene graphs for open-world object-goal navigation. *The International Journal of Robotics Research*, page 02783649251369549.

[33] Rishabh Madan, Rajat Kumar Jenamani, Vy Thuy Nguyen, Ahmed Moustafa, Xuefeng Hu, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Sparcs: Structuring physically assistive robotics for caregiving with stakeholders-in-the-loop. *IROS*, 2022.

[34] Rishabh Madan, Skyler Valdez, Kim David, Fang Sujie, Zhong Luoyan, Virtue Diego, and Tapomayukh Bhattacharjee. Rabbit: A robot-assisted bed bathing system with multimodal perception and integrated compliance. *HRI*, 2024.

[35] Daniel Morton and Marco Pavone. Safe, task-consistent manipulation with operational space control barrier functions. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 187–194, 2025. doi: 10.1109/IROS60139.2025.11246389.

[36] Amal Nanavati, Ethan K. Gordon, Taylor A. Kessler Faulkner, Yuxin (Ray) Song, Jonathan Ko, Tyler Schrenk, Vy Nguyen, Bernie Hao Zhu, Haya Bolotski, Atharva Kashyap, Sriram Kutty, Raida Karim, Liander Rainbolt, Rosario Scalise, Hanjun Song, Ramon Qu, Maya Cakmak, and Siddhartha S. Srinivasa. Lessons learned from designing and evaluating a robot-assisted feeding system for out-of-lab use. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '25, page 696–707. IEEE Press, 2025.

[37] Abby O'Neill et al. Open X-Embodiment: Robotic learning datasets and RT-X models, 2023.

[38] Jordi Palacín, Eduard Clotet, Dani Martínez, David Martínez, and Javier Moreno. Extending the application of an assistant personal robot as a walk-helper tool. *Robotics*, 8(2):27, 2019.

[39] Daehyung Park, Yuuna Hoshi, Harshal P. Mahajan, Ho Keun Kim, Zackory Erickson, Wendy A. Rogers, and Charles C. Kemp. Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned. *Robotics and Autonomous Systems*, 124:103344, 2020. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2019.103344. URL https://www.sciencedirect.com/science/article/pii/S0921889018307061.

[40] Daehyung Park, Yuuna Hoshi, Harshal P Mahajan, Ho Keun Kim, Zackory Erickson, Wendy A Rogers, and Charles C Kemp. Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned. *Robotics and Autonomous Systems*, 124:103344, 2020.

[41] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2109.08238.

[42] Zachary Ravichandran, Alexander Robey, Vijay Kumar, George J Pappas, and Hamed Hassani. Safety guardrails for llm-enabled robots. *IEEE Robotics and Automation Letters*, 2026.

[43] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone. 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans. In *Robotics: Science and Systems (RSS)*, 2020.

[44] Zhe Su, Oliver Kroemer, Gerald E Loeb, Gaurav S Sukhatme, and Stefan Schaal. Learning manipulation graphs from demonstrations using multimodal sensory signals. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 2758–2765. IEEE, 2018.

[45] Zhanyi Sun, Yufei Wang, David Held, and Zackory Erickson. Force-constrained visual policy: Safe robot-assisted dressing via multi-modal sensing. *IEEE Robotics and Automation Letters*, PP:1–8, 05 2024. doi: 10.1109/LRA.2024.3375712.

[46] Katherine M Tsui, Sarah Cohen, Selma Sabanovic, Alex Alspach, Rune Baggett, David Crandall, and Steffi Paepcke. Perspective chapter: Uncovering older adult needs–applying user-centered research methodologies to inform robotics development and a call to action. 2024.

[47] Katherine M. Tsui, Rune Baggett, and Carol Chiang. Exploring embodiment form factors of a home-helper robot: Perspectives from care receivers and caregivers. *Applied Sciences*, 15(2), 2025. ISSN 2076-3417. doi: 10.3390/app15020891. URL https://www.mdpi.com/2076-3417/15/2/891.

[48] Weikang Wan, Yifeng Zhu, Rutav Shah, and Yuke Zhu. Lotus: Continual imitation learning for robot manipulation through unsupervised skill discovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 537–544. IEEE, 2024.

[49] Zhanxin Wu, Bo Ai, Tom Silver, and Tapomayukh Bhattacharjee. Savor: Skill affordance learning from visuo-haptic perception for robot-assisted bite acquisition. In *Conference on Robot Learning (CoRL)*, 2025.

[50] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *Robotics: Science and Systems XIX*, 2024.

[51] Ruolin Ye, Wenqiang Xu, Haoyuan Fu, Rajat Kumar Jenamani, Vy Nguyen, Cewu Lu, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Rcareworld: A human-centric simulation world for caregiving robots. *IROS*, 2022.

[52] Ruolin Ye, Yifei Hu, Yuhan Anjelica Bian, Luke Kulm, and Tapomayukh Bhattacharjee. Morpheus: a multimodal one-armed robot-assisted peeling system with human users in-the-loop. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9540–9547, 2024. doi: 10.1109/ICRA57147.2024.10610050.

[53] Uksang Yoo, Nathaniel Dennler, Eliot Xing, Maja Mataric, Stefanos Nikolaidis, Jeffrey Ichnowski, and Jean Oh. Soft and compliant contact-rich hair manipulation and care. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, 2025.

[54] Fan Zhang and Yiannis Demiris. Learning garment manipulation policies toward robot-assisted dressing. *Science robotics*, 7(65):eabm6010, 2022.

[55] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

[56] Chengyang Zhao, Uksang Yoo, Arkadeep Narayan Chaudhury, Giljoo Nam, Jonathan Francis, Jeffrey Ichnowski, and Jean Oh. Dymo-hair: Generalizable volumetric dynamics modeling for robot hair manipulation. *International Conference on Robotics and Automation*, 2026.

[57] Jian Zhao, Yunlong Lian, Andy Tyrrell, Michael Gienger, and Jihong Zhu. Bimanual robot-assisted dressing: A spherical coordinate-based strategy for tight-fitting garments. pages 3328–3335, 10 2025. doi: 10.1109/IROS60139.2025.11246012.

[58] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-

action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[59] Athanasia Zlatintsi, Isidoros Rodomagoulakis, Petros Koutras, AC Dometios, Vassilis Pitsikalis, Costas S Tzafestas, and Petros Maragos. Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 3171–3175. IEEE, 2018.