Imitation Learning through Action Understanding: A Review on Robotic Manipulation

Milad Rabiei *

***Abstract*—Recent advances in robotic manipulation highlight a growing emphasis on leveraging vision-based action segmentation to bridge human demonstrations and autonomous robotic planning. This body of research reflects a shift from shallow action recognition toward rich, hierarchical models that capture the temporal and semantic structure of human activities. Central to this evolution is the extraction of symbolic and subsymbolic action primitives from multimodal data—visual, kinematic, and contextual—which enables more flexible and cognitively informed planning in robots. Techniques such as learning from demonstration, procedural task modeling, and diffusion-based planning are increasingly integrated to allow robots to generalize across tasks and adapt to real-world uncertainties. Despite these strides, the field continues to grapple with challenges in domain transfer, robustness to noisy demonstrations, and autonomy in unstructured environments. Addressing these limitations remains essential for the deployment of robots capable of understanding, imitating, and reasoning about human intent in dynamic, collaborative settings.**

***Index Terms*—Robotic manipulation, action segmentation, learning from demonstration, symbolic and subsymbolic representation, vision-based planning, task primitives, human-robot collaboration, cognitive robotics.**

I. Introduction

Robotic object manipulation remains one of the most challenging and consequential areas of intelligent robotics, particularly when applied to unstructured environments that require fine-grained adaptability, contextual awareness, and generalization. As robots increasingly enter human-centric domains—homes, hospitals, warehouses, and collaborative workspaces—the need for systems capable of understanding and replicating human actions with dexterity and reliability becomes paramount. Human demonstrations offer a powerful learning signal for this goal, encapsulating rich information about task structure, intent, and environmental interaction. However, leveraging this data for robotic learning requires robust mechanisms for action segmentation: the ability to break down complex manipulation sequences into meaningful, reusable components that machines can interpret and execute.

This systematic review investigates how contemporary vision-based methods contribute to action segmentation in robotic object manipulation from human demonstrations. It asks how recent research enables the extraction, representation, and transfer of manipulation knowledge, with particular attention to the challenges of learning from diverse, ambiguous, or noisy real-world demonstrations. The review is structured around four core research questions that collectively address the challenges of skill acquisition, spatio-temporal representation, intent inference, and sim-to-real generalization.

* Dipartimento di informatica, bioingegneria, robotica e ingegneria dei sistemi - DIBRIS, Università degli Studi di Genova, Genoa, Italy

A. Significance

The significance of this topic lies in its potential to advance both the theoretical foundations and practical capabilities of robot learning. Action segmentation serves as a critical interface between low-level sensory input and high-level planning and control. By identifying and formalizing the building blocks of human behavior—through primitives, keystates, or symbolic abstractions—robots can move beyond naive imitation toward intelligent, context-aware action generation. Moreover, by integrating insights from temporal logic, causal modeling, and hierarchical learning, researchers are progressively enabling robotic systems to understand not only what actions to perform, but also why and when to perform them.

In summarizing and analyzing the current landscape, this review contributes to a deeper understanding of the representational choices, algorithmic strategies, and data assumptions that shape the field. It highlights promising directions for future research, such as multi-modal representation learning, symbolic-subsymbolic integration, and scalable learning from unstructured video. Ultimately, the insights derived from this review serve to inform the development of more intelligent, adaptable, and general-purpose robotic agents capable of functioning in dynamic human environments.

B. Research Questions

This systematic review investigates how recent research contributes to advancing action segmentation for robotic object manipulation based on human demonstrations. Specifically, how do state-of-the-art vision-based methods perform action segmentation in the context of robotic object manipulation, and in what ways do they enable the extraction, representation, and transfer of human demonstration knowledge to robotic systems? To guide this inquiry, the review explores several questions:

1. How can robots effectively learn complex manipulation skills from human demonstrations?

2. What representational frameworks are most suitable for capturing the spatio-temporal aspects of human manipulation for robot learning and planning?

3. How can human intent, causality, and knowledge be extracted from human demonstrations to enable more adaptable robot behavior?

4. How can robots effectively bridge the sim-to-real gap and generalize over policies to real-world environments?

By answering these questions, the review aims to synthesize trends, frameworks, and methodological innovations that collectively define the current landscape of the field, and to identify directions for future research.

C. Scope

This systematic review focuses on recent methodological contributions to action segmentation in the context of robotic object manipulation. The literature search was conducted using Google Scholar, employing the terms "action segmentation" and "domain" and "manipulator" and "IEEE" to identify relevant studies. To ensure the inclusion of contemporary work, the review was limited to publications published in the years 2020 to 2025, written in English, and sourced exclusively from IEEE journals, letters, and conference proceedings.

Following an initial screening of 474 search results, a set of exclusion criteria was applied to refine the corpus. Gaze-based systems were excluded, as the review centers on object manipulation rather than user-attention or eye-tracking-based interaction paradigms. Studies that did not address object manipulation tasks—such as those focusing on deformable object handling or tool use unrelated to object-oriented manipulation—were also excluded. Furthermore, only systems relying solely on visual data were considered; works involving tactile, auditory, or multi-modal sensing were omitted to maintain consistency in sensory modality. Papers centered primarily on datasets or benchmarking, rather than methodological advancements in action segmentation, were excluded. In addition, studies that dealt exclusively with low-level motion primitives, such as joint-level trajectory segmentation, were not included, as the review emphasizes higher-level task abstractions in action analysis. Citation metrics were deliberately disregarded during the selection process to avoid bias toward older publications and to ensure a focus on current research trends. Finally, 14 papers, listed concisely in Table 1, were selected for this review.

This review is organized thematically around four core research questions on action segmentation for robotic object manipulation from human demonstrations. Section II provides a concise introduction to some of the concepts discussed in the review. Section III summarizes the literature taken into account. Section IV analyzes the methods described in the previous section by discussing open-still challenges and research gaps, and tries to provide answers to the main research questions. Finally, section V concludes the review.

II. Brief Summary and Overview

In robotics, a task typically refers to a predefined goal or activity that a robot is expected to complete. These tasks often consist of sequences of atomic actions that require interaction with both the robot's environment and its internal state, involving planning, perception, and physical execution. The sources reviewed for this study provide important insight into the challenges and methodologies for such tasks.

Robot manipulation is an important part of human-robot interaction technology. In-hand manipulation is a specialized subdomain, focusing on a robot's ability to control and reposition objects withing the hand without external regrasping. The level of difficulty may be affected by the level of dexterity of the robot itself. A robust strategy is necessary to ensure that the final grasp after manipulation is functionally valid; An incorrect grasp can result in failure of subsequent task steps-such as obstructing an insertion path or compressing tool effectiveness. Besides from proper state transition, planning is also vital for any sort of manipulation. Informally, planning and control answer the questions of what and how, respectively. Planning-typically formulated using symbolic methods-must be reliable and safe for real-world implementation viability. While symbolic planning methods provide modularity and interpretability (and subsequently reliability) they require significant manual effort to design domains and constraints. To address this bottlenech, many recent approaches favor reinforcement learning (RL) and deep reinforcement learning (DRL), which allow agents to learn optimal behavior through interaction with the environment. Apricot [1] addresses important challenges in DRL, sparsity of rewards and sample efficiency, for in-hand tool manipulation. The authors propose action primitives based on contact-state transition, to decompose the manipulation into short-term action primitives: detach, crossover, and attach. While this approach does not directly solve action segmentation, it suggests how breaking raw data into low-dimensional information eases the problem solving of a RL model. By training a policy for each primitive, the learning process becomes more tractable. The contact-state is defined based on finger arrangement, grasp stability, and manipulability. The approach also suggests generalization by randomization of object shapes, friction coefficients, and observed joint angles. The framework highlights the potential for reusable primitives to achieve manipulation operations.

In the from human to robot everyday activity [3] paper, the goal is to enhance robotic cognitive reasoning by identifying, collecting, and describing human behaviors in everyday activities. This paper achieves success using multi-modal data, which puts it outside the scope of this review. However, it adds to the validity of learning from demonstration for manipulation through vast amounts of data.

From human-human collaboration to human-robot collaboration [2] proposes a system for automatically constructing task knowledge models from dual-human demonstrations in a real environment, an automation that clears the need for the time-consuming nature of defining knowledge models, task graphs, and skill libraries. The system first segments video demonstrations into sequences of action primitives using a vision-based parser and heuristic rules. It then employs a graph-based algorithm to extract task structure information, producing task graphs. Finally, action primitives, interactive information between agents (e.g., handover actions), and temporal constraints are modeled into a structured semantic model, which serves as a robot skill library with query and reasoning interfaces. The proposed methods were validated in an IKEA table assembly task, demonstrating that a robot could learn to collaborate as an assistant by observing human-human interactions.

Action primitives stand higher in the abstraction hierarchy than kinematic primitives. Kinematic primitives serve as fundamental (and low-level) motion units, often discovered by segmenting velocity or acceleration profiles over time. The paper hierarchical segmentation of human manipulation movements [4] introduces a hierarchical segmentation algorithm designed to split complex human manipulation movements

TABLE I
SUMMARY OF REVIEWED PAPERS

	Title	Year	Tags and Properties	Brief Summary
1	APriCoT: Action Primitives Based on Contact-State Transition [1]	2024	In-hand manipulation, contact transitions, action primitives, DRL, IsaacGym	Proposes three abstract action types (detach, crossover, attach) to model contact-state transitions for in-hand tool manipulation.
2	From Human-Human to Human-Robot Collaboration [2]	2022	Task graph modeling, vision-based parser, LfD, dual-agent demo	Extracts action primitives and task structure from dual-human demos using vision and heuristic segmentation.
3	From Human to Robot Everyday Activity [3]	2020	Multimodal data, ontological reasoning, activity recognition, EEG	Presents a data processing pipeline for activity classification, not segmentation; integrates vision, text, and EEG.
4	Hierarchical Segmentation of Human Manipulation Movements [4]	2022	Hierarchical action segmentation, velocity-based change detection	Splits complex human actions into movement primitives and grouped actions using supervised and unsupervised methods.
5	Keystate-Driven Long-Term Generation [5]	2025	Action keystates, bimanual generation, motion dictionary, LLMs	Uses LLMs and motion dictionaries to forecast long bimanual manipulation sequences via semantic keystates.
6	Learning Dictionaries of Kinematic Primitives [6]	2021	Unsupervised, kinematic motion primitives, velocity minima	Builds a primitive dictionary from video by segmenting velocity profiles to enable action classification.
7	Learning Symbolic and Subsymbolic Temporal Constraints [7]	2024	Symbolic/subsymbolic modeling, bimanual demos, GMMs	Infers symbolic temporal constraints from timing differences between semantic keypoints in demonstrations.
8	Learning to Visually Connect Actions and Their Effects (CATE) [8]	2024	Action-effect causality, self-supervised video representation	Introduces Action Selection and Effect-Affinity tasks for understanding causal relations between actions and outcomes.
9	LTLDoG: Symbolic Constraints for Diffusion-Based Planning [9]	2024	LTLf, diffusion models, symbolic planning, safety constraints	Uses LTLf logic to guide diffusion-based planners in generating temporally and symbolically constrained trajectories.
10	Motion Reasoning for Goal-Based Imitation Learning [10]	2020	Goal inference, trajectory-based intent reasoning, object motion	Uses object trajectories for action segmentation to infer whether observed actions are goal-directed or incidental.
11	Multi-Task Learning of Object States and Actions [11]	2024	Web video, self-supervised, object state transition	Trains a multitask model to detect object state changes and associated actions using noisy internet videos.
12	PlaTe: Visually Grounded Planning With Transformers [12]	2022	Transformer planning, procedural tasks, visual latent space	Uses a Transformer to model decision-making and latent state-action spaces without explicit segmentation.
13	Robotic Imitation of Human Actions [13]	2024	Diffusion-based segmentation, open-vocab object detection	Leverages diffusion models and object detection to segment and imitate actions from single video demonstrations.
14	Sim-to-Real Domain Shift in Online Action Detection [14]	2024	Sim-to-real, dataset contribution, temporal variability	Focuses on sim-to-real generalization for action detection and introduces a dataset highlighting action complexity.

into basic movement building blocks and then merge them forming actions. The segmentation performs a velocity-based inference algorithm, which identifies movements characterized by a bell-shaped velocity profile. The algorithm combines unsupervised segmentation with a supervised classification of these building blocks into labeled actions, and also offers an unsupervised clustering variant. Evaluations on both one-handed point-to-point movements and dual-arm object rotation tasks demonstrated that the approach reliably detects both building blocks and actions, even with small labeled training datasets. The clustering-based variant showed superior performance for dual-arm movements, highlighting its robustness without requiring extensive manual training data.

Perception and data representation are also crucial for manipulation planning and learning. Systems often leverage pose estimation, object modeling, and motion capture data to understand and replicate human demonstrations. Humanobject interaction data—annotated with key poses and labeled actions—serve as rich sources for training imitation learning models. Keystate-driven long-term generation of bimanual object manipulation sequences [5] presents a unified framework for the long-term generation of bimanual object manipulation sequences, aiming to overcome limitations of previous methods, such as prediction degradation over long durations, averaging out of fine motions, and unrealistic hand-object contact. The core innovation is to decompose long sequences into shorter subsequences defined by keystates, which mark the semantic start or end of an action. A large language model (LLM) is used to anticipate future action keystate labels, followed by a network that generates the corresponding keyposes, and another that fills in the motion between these keyposes. For fine-grained periodic motions (like cutting or stirring), a motion dictionary is introduced to store and retrieve raw trajectories, preventing their averaging out by MSE-based regression.

Imitation Learning (IL), a subclass of RL, bypasses the need for explicit reward engineering by training agents to mimic expert demonstrations. It is often employed for both planning and control, with trajectory planning emerging as a key application area. Furthermore, diffusion models have recently gained traction for long-horizon trajectory generation, offering generative capabilities for smooth and temporally

consistent action sequences. Such approaches highlight the importance of action primitives sequential and precedence rules. One method proposes a planning framework that integrates symbolic constraints into a diffusion-based trajectory generation process. By using a differentiable function derived from temporal logic constraints, the model generates safe and goal-aligned trajectories, generalizing to novel instructions and environments during deployment. It demonstrates applicability to both navigation and manipulation tasks in simulation and on real robots [9].

Another approach focuses on learning symbolic and subsymbolic temporal constraints from human bimanual demonstrations. By modeling the timing relationships between actions through keypoint differences and using GMMs and fuzzy logic, the method captures precise and flexible temporal coordination required for complex dual-arm tasks [7]. Such constraints help define knowledge models, and develop symbolic plans later on.

Motion dictionaries can be used as memory modules to retrieve or generate contextually appropriate motion segments. These may also support the generation of realistic interaction dynamics, essential for physical plausibility in execution. A kinematics-based method segments human motion into primitives using unsupervised learning on velocity profiles. These are clustered into a dictionary of atomic movements that serve as basic units for action classification. The representation shows robustness across viewpoints and provides a compact motion vocabulary for downstream robot planning and recognition tasks [6]. The approach demonstrated tolerance to viewpoint changes, supporting cross-view action recognition. The simplicity and kinematic focus of the method make it a potential backbone for general action understanding.

A video understanding framework learns to associate actions with their visual consequences in a self-supervised manner. It introduces two tasks: selecting appropriate actions to produce a visual change (state transition), and estimating how directly an effect resulted from an action. These insights are foundational for learning intuitive dynamics without labels, supporting generalization in robotic agents [8]. Quantitative results showed that self-supervised CATE pre-training outperformed other state-of-the-art methods in video representation learning and action quality assessment. The study highlights the fundamental role of understanding action dynamics and their effects for autonomous agents.

Another method uses motion reasoning to infer the goals (intent) behind demonstrated actions. By reasoning about low-level trajectories rather than explicit action labels, it distinguishes between intentional and incidental movements. This inverse planning approach allows robots to correctly reproduce tasks even when human demonstrations include irrelevant motions [10]. The goal-based imitation from observation setup enables robots to reproduce tasks extracted from demonstrations in different environments. Experiments on a kitchen dataset demonstrated that motion reasoning significantly improved the success rate of goal recognition and enabled a robot to successfully reproduce tasks in a real kitchen, even when the demonstrator's actions involved objects that the robot did not need to manipulate in its own environment.

To enable large-scale learning of manipulation skills, one paper introduces a self-supervised method to extract object states and state-modifying actions from web videos. It leverages the natural causal and sequential structure of interactions, using multi-task learning to generalize and demonstrating zero-shot capabilities and robustness to noisy data [11].

A Transformer-based planning model learns structured latent representations of procedural tasks from instructional videos. By maximizing the likelihood of action sequences conditioned on visual start and goal states, the model enables visually grounded long-horizon planning. The learned latent space is shown to support real-world manipulation execution. [12]. Furthermore, the possibility of applying learned procedural tasks was validated on a real UR-5 robot arm, showing its practical applicability for robotic manipulation.

Another work [13] presents a one-shot imitation framework that combines diffusion-based action segmentation, focusing on fundamental actions like grasping, holding, and manipulating, and open-vocabulary object detection to extract symbolic plans from human demonstrations. This spatial and temporal knowledge is then refined using symbolic reasoning to create an action plan, which is executed by the robot using inverse kinematics that account for its specific body schema.

A study on sim-to-real transfer emphasizes the use of synthetic video data to improve online action detection [14]. It introduces a benchmark with paired real and virtual egocentric videos and shows that synthetic data can bridge domain gaps, enhancing generalization to unseen actions while mitigating data collection burdens.

III. Critical Analysis and Research Gaps

several common methodological limitations and research gaps emerge, particularly concerning the transition from theoretical models and simulated environments to practical, robust robot deployment in the real world.

1) Domain Generalization and Real-World Noise: Many studies lack real-world grounding and confidently rely on simulation results, namely Apricot [1] and subsymbolic [7] papers. The main challenge regarding this matter is however running in a highly simplified and controlled environment. [2] validates its system on a simple IKEA table without additional scene objects, Plate [12] tests a relatively simple action, and motion reasoning [10] uses a mockup kitchen environment. Some methods additionally revolve around specially prepared data. [4] uses marker-based motion tracking, which is not best for the unconstrained in-the-wild videos. Keystate [5] also relies on a manually created look-up table for motion retrieval.

While Multi-Task Learning [11] leverages noisy uncurated web videos, it still notes that a significant portion of these videos are not related to the interaction category of interest, and their model struggles with simple negative sampling strategies. This indicates challenges in learning from uncontrolled, in-the-wild data without some level of curation or sophisticated noise-filtering mechanisms. The accuracy of vision-based methods can be significantly impacted by occlusions. Human-Human to Human-Robot [2] reports failures due to misestimation of visual tracking where hands may be blocked

by objects. This highlights a persistent challenge in real-world computer vision for robotics. Models can still fail when encountering scenarios not seen during training. Keystate-Driven [5] shows failures when a knife present instead of a peeler for a peel motion, or when an object is placed on the floor, an out-of-distribution scenario.

2) Limitation in Action Understanding: Most papers identify the challenge of inherent ambiguity in human motions. Motion Reasoning [10] argues that goals can be ambiguous at the symbolic action level and requires reasoning at the "low-level motion trajectories" to infer intent. CATE [8] identifies the "significant performance gap" between humans and machines in connecting actions and their effects. Keystate-Driven [5] also highlights "Action Ambiguity" (same initial action, multiple purposes) as a unique challenge not fully resolved. Some methods rely on averaging state variables over a demonstration, losing detail of fine-grained motions and subtle state changes, like cutting or stirring. Those are flattened due to near-static behavior around higher-velocity profile movements.

Further, some utilize predefined structures for actions. APri-CoT [1] acknowledges the labor-intensive nature of manually designing the graph for all operations. Similarly, Human-Human to Human-Robot [2] relies on heuristic rules based on human knowledge for action segmentation, rather than fully autonomous discovery. The action spaces used are often simplified or pre-defined, limiting generalizability. Motion Reasoning [10] focuses on standard pick-and-place plus pour and cook. Multi-Task Learning [11] restricts itself to irreversible actions to simplify the learning problem, missing a large class of everyday tasks.

3) Scalability: Diffusion models, as used in LTLDoG [9], usually require significant amounts of training data and many diffusion steps during inference, posing a practical challenge. Keystate-Driven [5] addresses computational cost by segmenting tasks, but LLM reliance can still be a factor.

A. Connect to Research Questions

The reviewed body of work significantly advances the field of robot learning from human demonstrations by addressing several foundational research questions central to human-robot interaction. One key area of focus is enabling robots to learn complex and dexterous manipulation skills, particularly in tasks involving dynamic object interactions. Methods like APriCoT [1] propose decomposing manipulation into action primitives based on contact-state transitions, allowing complex actions to be structured into simpler, learnable units. Keystate-Driven [5] takes a different approach by identifying key transitional states within bimanual manipulation sequences and using a motion dictionary to generate precise actions such as cutting and stirring. Similarly, the Kinematic Primitives [6] framework breaks down demonstrations into reusable sub-movements, providing a compact representation of fundamental manipulation behaviors.

Another important contribution of the literature lies in exploring representational frameworks that effectively capture the spatial, temporal, and causal properties of manipulation. Approaches that integrate symbolic and subsymbolic representations—such as Symbolic and Subsymbolic [7]—highlight the value of combining temporal logic with sensorimotor features to form robust models of human activity. Hierarchical segmentation methods like [4] further contribute by organizing motion data into multi-level abstractions, enhancing interpretability and reuse. The use of formal temporal logic in LTLDoG [9], combined with diffusion models, demonstrates how high-level human instructions can be encoded and satisfied through trajectory generation, connecting symbolic planning and continuous motion execution.

The literature also tackles the challenge of understanding implicit human intent and causality from diverse demonstrations. Motion Reasoning [10], for instance, emphasizes distinguishing between intentional goals and incidental movements using low-level object interactions. CATE [8] proposes a novel task structure to directly assess the causal relationship between actions and visual effects, while Multi-Task Learning [11] uses uncurated web videos as a scalable source of causal action-effect pairs by leveraging temporal ordering cues. Meanwhile, EASE [3] enriches this line of inquiry by incorporating multimodal data—including biological signals and verbalized thought processes—to gain insight into the cognitive dimensions of human activity.

Finally, the question of sim-to-real transfer and policy generalization remains a persistent challenge across the studies. Many works, such as Sim-to-Real Domain Shift in Online Action Detection [14], directly investigate this gap by evaluating how synthetic data can improve recognition and prediction in real-world scenarios. PlaTe [12] demonstrates a promising direction by employing Transformer-based models to learn plannable latent spaces from visual demonstrations, validating on a physical robot in simplified tasks. Collectively, these contributions highlight the field's efforts to move beyond controlled environments and develop models that are both generalizable and robust to real-world noise, occlusion, and variability.

IV. Conclusion

The body of work presented in these sources demonstrates a robust and evolving landscape in the field of robotic manipulation, particularly concerning the use of vision for action segmentation and its subsequent role in domain generation for manipulator planning. Researchers are moving beyond simple action recognition to create sophisticated, hierarchical representations of human activities, incorporating both symbolic and subsymbolic cues. The decomposition of complex tasks into manageable primitives, the integration of diverse visual and kinematic data, and the application of advanced machine learning techniques are key enablers. The ultimate goal is to equip robots with the cognitive ability to interpret human intent, plan complex actions, and adapt to varied real-world scenarios, bridging the gap between human demonstrations and autonomous robotic execution. At the same time, the literature reveals persistent challenges related to data quality, generalization, and system autonomy—critical issues that future research must address to enable reliable robot deployment in dynamic, unstructured human environments.

REFERENCES

[1] D. Saito, A. Kanehira, K. Sasabuchi, N. Wake, J. Takamatsu, H. Koike, and K. Ikeuchi, "APriCoT: Action Primitives based on Contact-state Transition for In-Hand Tool Manipulation," arXiv preprint arXiv:2407.11436, 2024. [Online]. Available: https://arxiv.org/abs/2407.11436

[2] Y. You, Z. Ji, X. Yang, and Y. Liu, "From human-human collaboration to human-robot collaboration: automated generation of assembly task knowledge model," in *Proc. 27th Int. Conf. Automation and Computing (ICAC)*, Bristol, U.K., 2022, pp. 1–6. doi: 10.1109/ICAC55051.2022.9911131.

[3] C. Mason *et al.*, "From Human to Robot Everyday Activity," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2020, pp. 8997–9004. doi: 10.1109/IROS45743.2020.9340706.

[4] L. Gutzeit, "Hierarchical Segmentation of Human Manipulation Movements," in *Proc. 26th Int. Conf. Pattern Recognition (ICPR)*, 2022, pp. 2742–2748. doi: 10.1109/ICPR56361.2022.9955634.

[5] H. Razali and Y. Demiris, "Keystate-Driven Long-Term Generation of Bimanual Object Manipulation Sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 01, pp. 1–12, May 5555. doi: 10.1109/TPAMI.2025.3573081.

[6] A. Vignolo, N. Noceti, A. Sciutti, F. Odone, and G. Sandini, "Learning dictionaries of kinematic primitives for action classification," in *Proc. 25th Int. Conf. Pattern Recognition (ICPR)*, 2021, pp. 5965–5972. doi: 10.1109/ICPR48806.2021.9412363.

[7] C. Dreher and T. Asfour, "Learning Symbolic and Subsymbolic Temporal Task Constraints from Bimanual Human Demonstrations," arXiv preprint arXiv:2403.16953, 2024. [Online]. Available: https://arxiv.org/abs/2403.16953

[8] P. Parmar, E. Peh, and B. Fernando, "Learning to Visually Connect Actions and their Effects," arXiv preprint arXiv:2401.10805, 2024. [Online]. Available: https://arxiv.org/abs/2401.10805

[9] Z. Feng, H. Luan, P. Goyal, and H. Soh, "LTLDoG: Satisfying Temporally-Extended Symbolic Constraints for Safe Diffusion-Based Planning," *IEEE Robot. Autom. Lett.*, vol. 9, no. 10, pp. 8571–8578, 2024. doi: 10.1109/LRA.2024.3443501.

[10] D.-A. Huang, Y.-W. Chao, C. Paxton, X. Deng, L. Fei-Fei, J. C. Niebles, A. Garg, and D. Fox, "Motion Reasoning for Goal-Based Imitation Learning," arXiv preprint arXiv:1911.05864, 2020. [Online]. Available: https://arxiv.org/abs/1911.05864

[11] T. Souček, J.-B. Alayrac, A. Miech, I. Laptev, and J. Sivic, "Multi-Task Learning of Object States and State-Modifying Actions From Web Videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5114–5130, 2024. doi: 10.1109/TPAMI.2024.3362288.

[12] J. Sun, D.-A. Huang, B. Lu, Y.-H. Liu, B. Zhou, and A. Garg, "PlaTe: Visually-Grounded Planning With Transformers in Procedural Tasks," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4924–4930, 2022. doi: 10.1109/LRA.2022.3150855.

[13] J. Spisak, M. Kerzel, and S. Wermter, "Robotic Imitation of Human Actions," arXiv preprint arXiv:2401.08381, 2024. [Online]. Available: https://arxiv.org/abs/2401.08381

[14] C. Patsch, W. Torjmene, M. Zakour, Y. Wu, D. Salihu, and E. Steinbach, "Sim-to-Real Domain Shift in Online Action Detection," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, Oct. 2024, pp. 388–394. doi: 10.1109/IROS58592.2024.10802421.