IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 September 2025, accepted 19 October 2025, date of publication 23 October 2025, date of current version 30 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3624723

SURVEY

Learning by Watching: A Review of Video-Based Learning Approaches for Robot Manipulation

CHRISANTUS EZE, (Graduate Student Member, IEEE), AND CHRISTOPHER CRICK, (Member, IEEE)
Computer Science Department, Oklahoma State University, Stillwater, OK 74078, USA
Corresponding author: Chrisantus Eze (chrisantus.eze@okstate.edu)

ABSTRACT Robot learning of manipulation skills is hindered by the scarcity of diverse, unbiased datasets. While curated datasets can help, challenges remain regarding generalizability and real-world transfer. Meanwhile, large-scale 'in-the-wild' video datasets have driven progress in computer vision using self-supervised techniques. Translating this to robotics, recent works have explored learning manipulation skills using abundant passive videos sourced online. Showing promising results, such video-based learning paradigms provide scalable supervision and reduce dataset bias. This survey reviews foundations such as video feature representation learning techniques, object affordance understanding, 3D hand and body modeling, and large-scale robotic resources, as well as emerging techniques for acquiring robot manipulation skills from uncontrolled video demonstrations. We discuss how learning from only observing large-scale human videos can enhance generalization and sample efficiency for robotic manipulation. The survey summarizes video-based learning approaches, analyzes their benefits over standard datasets, survey metrics and benchmarks, and discusses open challenges and future directions in this nascent domain at the intersection of computer vision, natural language processing, and robot learning.

INDEX TERMS Video, watching, robot manipulation, demonstration, imitation, reinforcement learning.

I. INTRODUCTION

In contrast to fields like computer vision (CV) and natural language processing (NLP), where copious amounts of high-quality and diverse datasets are available, the field of robotics faces a significant limitation in the availability of such datasets for various tasks. This scarcity of quality data has hindered progress in robotics in multiple ways. To address this challenge, researchers have proposed algorithms based on techniques like few-shot learning [1], [2], [3] and multitask learning [4], [5]. While these approaches show promise in mitigating the data scarcity issue, they still rely on a substantial amount of high-quality data for effective task generalization.

Similarly, classical robot planning and manipulation methods often necessitate detailed modeling of the world and agent dynamics, further limiting their transferability and generalizability. Despite efforts to employ deep reinforcement learning (RL) for motion planning [6], [7] and manipulation [8], [9], these methods encounter challenges such as distribution shifts and reductions in generalizability. Recent imitation learning methods [10], [11], [12] based on Behavioral Cloning (BC) [13] have also emerged as a potential solution to learning manipulation skills from minimal demonstrations. However, similar to their deep RL counterparts, these methods struggle to learn manipulation skills in diverse and uncurated datasets.

In recent times, significant strides have been made in collating large-scale, high-quality datasets for diverse robotic tasks [14], [15], [16], [17], [18], [19] akin to the impact of the ImageNet dataset in the field of Computer Vision [20]. While this marks a positive step forward, these datasets often exhibit limitations in their representativeness of real-world environments, as they are typically collected in controlled settings. Despite their advantages, these datasets pose potential drawbacks, including limited generalizability, biases [21], high costs, and ethical concerns regarding the interactions of embodied agents with humans.

The associate editor coordinating the review of this manuscript and approving it for publication was Hongli Dong.

In contrast, ''in-the-wild'' datasets have played a pivotal role in the success of computer vision [22], [23], [24], [25], [26], particularly with the rise of self-supervised learning. In the realm of robotics, various works have embraced this approach, training embodied agents to acquire manipulation skills by learning from videos sourced from platforms like YouTube. These endeavors have demonstrated impressive performance improvements, showcasing enhanced generalizability.

This paper provides a comprehensive exploration of video-based learning methodologies, with a focus on addressing fundamental challenges in vision-based robotic manipulation. Specifically, we investigate the potential of these methodologies to enhance the learning of generalizable skills, mitigate biases, and reduce the costs associated with curating high-quality datasets. Our contributions are threefold: (1) a detailed review and analysis of the capabilities of current approaches in various robotic tasks, (2) an overview of some open-source resources and tools for video-based robot manipulation learning to help researchers get started, and (3) a discussion of current challenges and future directions in the field. Our work focuses solely on vision-based manipulation. We discuss navigation, locomotion, and non-visual-based approaches only to provide a more broad perspective. We commence by introducing and summarizing the foundational components of learning from videos, and then proceed to discuss current approaches for acquiring manipulation skills through video-based learning.

In Section II, we delineate and discuss the pipeline and essential components required for learning from video data. Additionally, we present notable large-scale robotic resources, including datasets and network architectures. Section III delves into the current approaches for learning from videos, categorized into five distinct groups, with a thorough literature review under each category. Section IV highlights the comparative analysis of the distinct categories of approaches. We present in Section V an overview of some open-source resources and tools used for video-based manipulation skill learning. Finally, Sections VI and VII summarize the existing challenges faced by researchers in developing systems for learning manipulation skills from videos, and propose potential research directions likely to have a significant impact in this domain.

A. SCOPE OF THIS SURVEY

This survey specifically explores techniques for acquiring robot manipulation skills through explicit learning from video data. Discussions on learning manipulation skills from data modalities other than videos or topics related to learning robot navigation skills are not within the scope of this article, though they may be briefly mentioned for a more comprehensive perspective. While foundational resources supporting this learning approach are touched upon, the primary focus is on discussing the techniques, advantages, and challenges associated with acquiring robust manipulation skills from video data. It is important to clarify that the list of foundational resources provided is not exhaustive and represents only the essential secondary components involved in the learning process. To the best of our knowledge, this survey is the first of its kind to explore the landscape of learning robot manipulation skills specifically from video data. Additionally, there is currently no existing work that comprehensively surveys the learning of robot skills in general, from videos, although many studies have surveyed the broader fields of robot manipulation and robot learning from demonstration.

B. RELATED SURVEYS

We examine survey articles already available on related subjects to guide readers to additional papers focusing on more specific topics. This serves the dual purpose of offering references for further exploration and elucidating the distinctions between this article and existing surveys.

In contrast to these prior surveys highlighted in Table 1, our work is uniquely focused on the intersection of video-based learning and robot manipulation. We systematically review and compare methods across both imitation and reinforcement learning that leverage video demonstrations as a core component, encompassing advances in vision-language models (VLMs), foundation models, and large-scale data resources. By specifically addressing challenges, generalization, and open questions in video-based manipulation learning, our survey fills a notable gap in the literature and provides a valuable reference for researchers seeking to navigate this rapidly evolving subfield.

We summarize and compare the specific areas of focus of these survey articles with ours in Table 1 below.

II. FOUNDATIONS OF LEARNING FROM VIDEOS

Learning robot manipulation skills from videos is a complex task that necessitates a comprehensive visual pipeline, encompassing various objectives such as representation learning, object affordance learning, human action recognition, and 3D hand modeling. In this section, we will delve into these objectives in detail and explore some of the proposed techniques for their execution.

A. REPRESENTATION LEARNING

Visual feature extraction forms the backbone of vision-based robotics. Over time diverse representation learning methods have emerged, each offering unique ways to capture meaningful features from visual data. These approaches broadly fall into two categories: those tailored for videos and those applicable to both images and videos.

Video-centric representation learning focuses on modeling temporal dynamics and multi-view consistency. For instance, Time-Contrastive Networks (TCN) [35] use self-supervised learning from multi-view videos to encode temporal changes while remaining invariant to viewpoint differences. Building on this idea, Domain-agnostic Video Discriminator (DVD) [36] employs multitask reward learning, training a discrim-

TABLE 1. A comparison of different aspects covered by existing surveys and our survey.

Paper	Cluttered Envs	LfD Techniques	RL	Generalized Algorithms	Foundation Models
[27]	✓	x	x	✓	x
[28]	x	✓	x	x	x
[29]	x	✓	x	x	x
[30]	x	x	✓	✓	x
[31]	x	x	✓	x	x
[32]	x	x	x	x	✓
[33]	x	x	x	x	✓
[34]	x	x	✓	x	✓
Ours	✓	✓	✓	x	✓

inator to verify whether two videos depict the same task, thereby extracting domain-invariant features.

Unsupervised approaches further extend these capabilities. Wang and Gupta [37] leverage visual tracking as supervision, aligning tracked patches across frames via a Siamese-triplet loss to learn rich representations from unlabelled web videos. While CNNs excel at spatial feature extraction, temporal modeling often benefits from sequence models. For example, [38] introduced an LSTM-based encoder-decoder for compact video representations, supporting sequence reconstruction and future prediction. Similarly, Dense Predictive Coding (DPC) [39] learns spatio-temporal embeddings in a self-supervised manner for tasks like action recognition.

Beyond spatial and temporal cues, some methods incorporate geometry and structure. Reference [40] proposed an unsupervised framework for jointly estimating monocular depth and camera motion using view synthesis as supervision, inspiring later work on 3D scene understanding [41], [42], [43]. Other methods, such as Contrastive Video Representation Learning (CVRL) [44], use contrastive learning to align augmented views while differentiating unrelated clips, producing robust spatiotemporal representations.

Moving toward general-purpose methods, recent research targets representations transferable across images and videos. Masked Modeling [45] showed that self-supervised pretraining on real-world images can outperform traditional ImageNet-based pretraining [20] in robotic manipulation benchmarks. Extending this idea, [46] applied masked autoencoders (MAE) to large-scale video data for visual pretraining, integrating these frozen representations into downstream control policies.

Finally, universal representations like R3M [47] combine time-contrastive learning with sparse encoding from human video datasets. Serving as a frozen perception module, R3M enables efficient imitation learning across both simulated and real-world robotic manipulation tasks.

Table 2 shows a concise summary of the comparison between the approaches discussed.

B. OBJECT AFFORDANCE AND HUMAN-OBJECT INTERACTION

A key step in enabling robots to acquire manipulation skills from videos is understanding object affordances: the actionable properties of objects, through the lens of human interaction. Over the years, research has progressed from early hand-state analysis in large-scale internet videos to increasingly sophisticated, multimodal, and context-aware frameworks. This subsection charts the evolution of these methods, highlighting their innovations and interconnections.

We begin with approaches that leverage large-scale, unstructured human activity videos to uncover affordances. For example, [48] extracts hand-state information from internet videos, laying a foundation for understanding human-object interaction at scale. Building on this, Hand-aided Affordance Grounding Network (HAG-Net) [49] employs hand cues from demonstration videos and a dual-branch network for fine-grained localization of affordance regions, improving the precision of affordance grounding.

The field then expands toward capturing functional understanding and temporal dynamics of affordances. The authors in [50] introduce a generative model that grounds object affordances by considering both spatial context and human intention, while a related approach [51] models objects and sub-activities as a Markov random field, addressing the challenge of acquiring descriptive labels for sub-activities and their corresponding affordances.

Learning from demonstration (LfD) videos is another significant direction. Demo2Vec [52] focuses on reasoning about object affordances using carefully curated demonstration videos, learning vector embeddings to predict interaction regions and support both human and robot understanding. In contrast, Vision-Robotics Bridge (VRB) [53] demonstrates how affordance models trained on diverse, in-the-wild internet videos can bridge the gap between human-centric video data and the requirements of robotic applications.

The integration of depth data and multi-modal inputs has further advanced affordance detection. AffordanceNet [54] exemplifies this by introducing an end-to-end deep learning method for identifying both objects and their affordances in RGB-D images, effectively handling multiclass affordance masks. Similarly, [55] reviewed the landscape of affordance detection methods, underscoring the importance of understanding the full range of object affordances for real-world robot intelligence. At the 3D level, [56] proposed semantic labeling of 3D point clouds to improve object segmentation and reduce uncertainty in manipulation, demonstrating that

TABLE 2. Comparison of representative video-based visual representation learning methods for robot manipulation.

Method	Task Performance	Sample Efficiency	Generalization	Compute Cost
[35]	Effective for third-person imitation; supports RL via embedding similarity	Learns from multi-view video; label-free but needs synchronization	Generalizes across views and agents with consistent context	Lightweight CNN + metric loss; training setup moderately complex
[36]	Solves real-robot tasks using learned video rewards	Requires only a few robot demos and 1 human video	Strong zero-shot generalization to new tasks and environments	Moderate; uses discriminators with broad human video data
[37]	Nearly matches ImageNet CNNs on VOC (52% mAP)	Uses 100K videos and patch tracking; no labels needed	Captures object-level similarity; weak on temporal cues	Siamese-triplet CNN; efficient but needs large-scale video processing
[38]	Predicts future video frames; helps with action recognition	Requires many sequences; less data-efficient	Weak on domain shift; struggles with unfamiliar dynamics	Moderate; encoder-decoder LSTMs for temporal modeling
[39]	75.7% on UCF101; strong for human action recognition	Avoids pixel prediction; uses curriculum for better efficiency	Robust to viewpoint and appearance variation	3D-ResNet + GRU; heavier than 2D CNNs but no reconstruction loss
[40]	Effective depth and ego-motion learning from video	Self-supervised via view synthesis; highly data-efficient	Generalizes across driving scenes; handles occlusions	Dual CNNs for depth and pose; warping increases training cost
[44]	70.4% on Kinetics-600; outperforms SimCLR and ImageNet	Leverages contrastive loss with smart augmentations	Temporal and spatial robustness; handles distant clip variations	R3D-50 backbone; moderate complexity with consistent training speed
[45]	Solves motor control tasks; up to 80% success	Pretrained on real-world images; no labels or robot data needed	Generalizes across tasks and robots	ViT-based MAE; costly pretraining, efficient inference
[46]	81% success in real-world robot tasks	Achieves strong results with only 20-80 demos per task	Generalizes across tasks, robots, and scenes	307M ViT encoder; high initial cost, efficient during deployment
[47]	+20% over MoCo/CLIP on 12 tasks; 50%+ success from 20 demos	Learns from human videos; effective with few demos	Strong task, view, and embodiment transfer	Sparse contrastive encoder; efficient for downstream control

incremental, multi-view merging can directly benefit manipulation planning.

The scope broadens to include human-object relationships and interaction recognition in video. The authors in [57] employed transformer architectures for joint spatial-temporal reasoning, while [58] leveraged hand localization in egocentric videos to understand object affordances. H20 dataset was introduced in [59], enabling synchronized multi-view RGB-D capture of two-handed object manipulation, providing rich annotations for developing and benchmarking new affordance-centric methods.

Moving from interaction recognition to grasp generation and segmentation, [60] highlights the consistency between hand contact points and object regions, introducing objectives for self-supervised training of grasp generation models. Reference [61] proposed a unified network to simultaneously predict 3D hand and object poses, model interactions, and recognize action categories in egocentric video sequences. Meanwhile, [62] developed a weakly supervised approach to segment hands and hand-held objects from motion in a single RGB image, leveraging motion-derived responsibility maps for network training.

Datasets play a pivotal role in advancing affordance learning. The authors in [63] introduced a comprehensive dataset with over 100K frames of hand-object interactions and rich 3D annotations. In parallel, [64] used computer vision to unify the identification of hand grip types, object properties, and action categories from images, providing a context-aware model of natural hand-object manipulation.

C. HUMAN ACTION AND ACTIVITY RECOGNITION

Recognizing human actions is essential for robots operating in human environments, particularly for understanding object affordances. Early work leveraged human action patterns from videos to guide robotic perception and decision-making [48], [49], [51], [65]. Building on this, [66] emphasized affordances over appearance, encoding object-hand interactions as strings to capture functional properties more robustly. To enable fine-grained recognition, [67] combined appearance and motion cues through convolutional networks, enhancing the discrimination of subtle action variations. For automated object interaction analysis, Interaction Region and Motion Trajectory prediction Network (IRMT-Net) [68] jointly estimates interaction regions and motion trajectories

from demonstrations, reducing reliance on manual guidance and improving adaptability across systems.

Recent advances include unsupervised and weakly supervised methods. Reference [69] proposed a framework that segments actions into sub-activities using alternating discriminative and generative learning, coupled with background modeling to filter irrelevant frames, achieving strong performance with minimal labeled data. Progress has also been driven by large-scale datasets such as UCF101 [70], which provides diverse, realistic user-uploaded videos for benchmarking action recognition algorithms.

D. 3D HAND MODELING

Bridging the embodiment gap between human hands and robot grippers is a key challenge in learning manipulation skills from videos. To address this, researchers have explored both hardware solutions and computational models for 2D/3D hand representation, enabling robots to more effectively imitate human actions.

Early approaches introduced anthropomorphic robotic hands for teleoperation and video-based learning. For example, LEAP Hand [71] is a low-cost design that supports visual teleoperation, passive learning, and sim-to-real transfer by extracting hand poses from web videos. Similarly, DexMV [11] provides a simulation and vision-based pipeline that maps 3D human hand poses to robot-compatible demonstrations, while DexVIP [72] leverages YouTube videos and human hand priors to learn dexterous grasping without expensive lab data, enabling generalization to novel objects.

Advances in motion capture have further improved fidelity. FrankMocap [73] offers fast monocular 3D hand and body pose estimation using SMPL-X [74], while MANO [75] provides a parametric hand model built on SMPL [76], delivering low-dimensional, realistic representations widely used in robotics and graphics. Complementary work [60] focuses on grasp realism, enforcing consistency between hand-object contact points through self-supervised objectives, improving flexibility and accuracy even during testing.

Collectively, these methods reduce embodiment differences and establish robust hand modeling pipelines, laying the foundation for high-fidelity manipulation learning from human demonstrations.

E. DATASETS

In our discussion of the essential components for training robot manipulation policies, the importance of datasets cannot be overstated. Datasets form the foundation upon which learning algorithms build their understanding of manipulation tasks. Learning robot manipulation skills from demonstration videos requires carefully curated datasets that capture humans performing these tasks in various environments such as kitchens, living rooms, workshops, and more. These datasets not only provide the raw data necessary for training but also offer insights into human-object interactions, task variability, and environmental context.

This section categorizes and details some of the most influential datasets used in the field of robot manipulation learning, highlighting their unique characteristics and contributions.

1) LARGE-SCALE VIDEO DATASETS

These datasets offer a vast amount of video data capturing diverse activities and interactions. Additionally, these datasets are typically sourced from the internet and present a wide range of scenarios and tasks, making them invaluable for generalizing robot learning.

- **YouTube**: As one of the largest video platforms, YouTube serves as a rich source of diverse video content. Several works have curated specific subsets of YouTube videos relevant to robots manipulation, providing a broad spectrum of tasks and environments.
  For instance, in [77], the authors introduced HD-VILA-100M, a large dataset with two distinct properties: 1) it is the first high-resolution dataset, including 371.5k hours of 720p videos, and 2) it is the most diversified dataset, covering 15 popular YouTube categories. YT-Temporal-180M, introduced in [78], is a diverse corpus of frames/ASR derived from a filtered set of 6M diverse YouTube videos.
  In addition to these works, researchers have introduced datasets with smoother and more descriptive video-text pairs. One of such works is WTS-70M, a 70M video clips dataset presented in [79], contains textual descriptions of the most important content in the video, such as the objects in the scene and the actions being performed. The authors in [80] introduced HowTo100M: a large-scale dataset of 136 million video clips sourced from 1.22M narrated instructional web videos depicting humans performing and describing over 23k different visual tasks. Additionally, [81] provided a new video-text pretraining dataset WebVid-10M, comprised of over two million videos with weak captions scraped from the internet.
- **Internvid**: Compiled from various internet sources, Internvid [82] focuses on activities and tasks that are particularly informative for robotic learning. This dataset encompasses a wide array of human activities, enhancing the versatility of trained models.
- **Something-Something**: This dataset [83] consists of videos where humans perform a wide range of actions on everyday objects. It is particularly useful for training models to recognize and replicate specific human-object interactions.

2) EGOCENTRIC (FIRST-PERSON) VIDEO DATASETS

Egocentric datasets capture videos from the first-person perspective, offering a unique vantage point for understanding hand-object interactions and human intent. These datasets are especially valuable for tasks that involve detailed manipulation and personal perspective.

- **Ego-4D**: A comprehensive dataset of first-person videos capturing daily activities, Ego-4D [84] provides rich data on hand-object interactions from the wearer's perspective. This dataset is instrumental in training models to understand and predict human actions in a personal context.
- **Ego-Exo-4D**: Building on the Ego-4D dataset, Ego-Exo-4D [85] includes both egocentric and exocentric (third-person) views of the same activities. This multi-perspective approach offers a more holistic understanding of tasks, aiding in the development of models that can interpret and execute actions from different viewpoints.

3) TASK-SPECIFIC AND MULTI-MODAL DATASETS

Task-specific and multi-modal datasets are designed to study particular tasks or provide multiple modalities of data, such as video, audio, and annotations. These datasets are tailored to enhance the learning process for specific manipulation skills.

- **Epic Kitchens**: Focused on kitchen activities, this dataset [86] captures detailed interactions with objects and the environment from an egocentric perspective. The rich annotations and diversity of tasks make it ideal for training models on kitchen-related manipulation tasks.
- **RoboVQA**: This dataset [87] is designed for Visual Question Answering in robotic contexts. It includes videos of robots performing tasks and corresponding questions that test the robot's understanding and reasoning based on the visual data. RoboVQA helps in developing models that can interpret and respond to queries about manipulation tasks.

4) EMBODIED AI AND INTERACTIVE DATASETS

Embodied AI and interactive datasets emphasize tasks that involve interaction with the environment, providing rich contextual information that is crucial for learning manipulation skills.

- **Open X-Embodiment**: A comprehensive dataset [16] that includes videos of various embodied AI tasks, capturing interactions in different environments. This dataset is the largest and most diverse open source robotics dataset to date, unifying 34 distinct datasets from 22 different robot embodiments. It is designed for large-scale, cross-platform model pretraining containing over 1.6 million trajectories spanning more than 60,000 unique tasks. It supports vision, language (instructions), and action (State/Pose) modalities.
- **DROID (Distributed Robot Interaction Dataset)**: DROID [88] comprises over 76,000 trajectories (∼350 hours) collected across 564 scenes and 86 tasks by more than 50 users. Designed for diversity and generalization, it outperforms Open X-Embodiment on both in- and out-of-distribution tasks, making it a strong benchmark for imitation learning.
- **BRMData (Bimanual-Mobile Robot Manipulation Dataset)**: BRMData [89] is a dataset that focuses on dual-arm and mobile manipulation, capturing tasks such as object handovers, opening cabinets, and cleaning. It features ten household tasks performed by a mobile manipulator equipped with two robot arms, and includes RGB and depth data from multiple camera viewpoints. The dataset also emphasizes environmental interaction and whole-body planning, supporting the development of controllers that operate in both tabletop and mobile contexts.
- **Fourier ActionNet**: Fourier ActionNet [90] contains 30,000 teleoperated bimanual trajectories (∼140 hours) of tabletop manipulation. Each trajectory is annotated with human-written task prompts, supporting instruction-conditioned policy learning and dexterous control.
- **Kaiwu Dataset**: Kaiwu [91] offers 11,664 demonstrations of human assembly tasks using 30 objects and 20 participants. The dataset includes synchronized RGB video, audio, EMG, eye gaze, motion capture, and tactile data, making it suitable for multimodal representation learning.
- **TASTE-Rob**: TASTE-Rob [92] provides over 100,000 egocentric video clips of human manipulation aligned with natural language instructions. It emphasizes object-centric motion and temporal segmentation, useful for training video-conditioned imitation policies.

F. LARGE SCALE ROBOTIC RESOURCES

The success of large-scale models in computer vision and natural language processing has set a high bar for what is possible with extensive data and powerful architectures. In robotics, a similar paradigm shift is underway with the introduction of Vision-Language-Action (VLA) models. The initial efforts in this space, represented by models like RT-1 [14] and RT-2 [15], provided the foundational blueprint for large-scale robot learning.

RT-1 was a major milestone, introducing open-ended, task-agnostic training and the Robotics Transformer architecture to enable strong generalization to new tasks with minimal data. The model uses a FiLM-conditioned [93] EfficientNet-B3 encoder, with instruction embeddings from a Universal Sentence Encoder, and compresses convolutional outputs using a TokenLearner module to produce compact visual tokens. These tokens, concatenated across the observation dimension, are fed into an 8-layer decoder-only transformer (∼19M parameters) that autoregressively predicts discretized action tokens, with each action dimension quantized into 256 bins. This design enabled real-time control and established the feasibility of training open-ended, task-agnostic robot policies with strong generalization. Building on this, RT-2 advanced the field by integrating VLMs. Instead of training a small transformer from scratch, RT-2 used pretrained VLMs such as PaLI-X [94] and PaLM-E [95]

TABLE 3. An overview of prominent video datasets relevant to robot learning. The table compares datasets on their primary content, unique features, scale, annotation methods, and data modalities.

Dataset	Content Focus	Key Feature	Scale	Modalities
General Web and Instructional Videos				
HowTo100M [80]	Instructional videos covering ~23k different human tasks	Massive scale for learning procedural, step-by-step tasks	136M clips from 1.22M videos (1.36M hours)	Vision, Language (ASR)
WebVid-10M [81]	Short, general-domain web videos with alt-text captions	Tightly-aligned, descriptive video-text pairs from web data	10.7M clips from 2.5M videos (52k hours)	Vision, Language (Alt-text)
Something-Something-v2 [83]	Basic human-object interactions (e.g., "pushing something")	Focus on fine-grained action recognition from templated labels	220,847 video clips	Vision, Language (Labels)
Egocentric Videos				
Ego-4D [84]	Daily life activities captured from a first-person view	Unprecedented scale and diversity for egocentric human activity	3,670 hours of video (2.78M clips)	Vision, Audio, Language, 3D Mesh, Eye Gaze, Stereo
Ego-Exo-4D [85]	Activities recorded simultaneously from ego- and exocentric views	Provides paired perspectives for learning view-invariant skills	1,400 hours of paired video	Vision, Audio, Language, Pose, 3D Geometry
EPIC-KITCHENS-100 [86]	Unscripted activities in a kitchen environment	Dense, fine-grained annotations of actions and object interactions	100 hours of video (90k action segments)	Vision, Audio, Language (Action/Object labels)
Robotics and Embodied AI				
Open X-Embodiment [16]	Manipulation trajectories from 22 different robot platforms	Unifies dozens of robotics datasets for large-scale co-training	1M+ trajectories across 527 skills	Vision, Language (Instructions), Action (State/Pose)
InternVid [82]	Narrated videos of diverse human-object interactions	High-quality, cleaned video-text pairs for strong generalization	236M clips from 7M videos (760k hours)	Vision, Language (Cleaned ASR)
RoboVQA [87]	Robot manipulation sequences for visual reasoning tasks	Designed for question-answering about robot actions and states	98k video-question pairs	Vision, Language (Q&A)
DROID [88]	Diverse manipulation demonstrations across many scenes and tasks	Collected by 50+ users globally with high generalization capability	76k trajectories (350 hours)	Vision, Language, Actions
BRMData [89]	Mobile and dual-arm household manipulation	Captures both tabletop and mobile dual-arm tasks in real homes	10 tasks with multi-view videos	Vision (RGB, Depth), Actions
Fourier ActionNet [90]	Bimanual dexterous manipulation	Teleoperated control with natural language task descriptions	30k trajectories (140 hours)	Vision, Language (Prompts), Actions
Kaiwu [91]	Human demonstrations of assembly tasks	Rich multimodal data including audio, gaze, EMG, and tactile sensors	11,664 demos across 30 objects	Vision, Audio, EMG, Eye Gaze, Tactile, Motion Capture
TASTE-Rob [92]	Egocentric hand-object manipulation	Large-scale video-instruction pairs for manipulation learning	100k+ video-instruction clips	Vision, Language (Instructions)

as perception backbones, casting robot actions as discrete language tokens. Actions (e.g., 7-DOF poses and gripper states) were serialized into integer strings, enabling the model to predict them like words in a sentence. These models, along with collaborative efforts like Open X-Embodiment [16], established the viability of training adaptable policies on diverse datasets collected from multiple robot platforms. However, their reliance on discretization for action modeling imposed limitations in precision and multimodality. This laid the groundwork for the next wave of research, which seeks to overcome these constraints through architectural innovations in action representation and reasoning.

The limitations of first-generation large models have sparked a wave of models [96], [97], [98], [99], [100], [101],

[102], [103] that prioritize more expressive action modeling and stronger reasoning capabilities. A central theme is the shift from simple action discretization to architectures that can capture continuity, multimodality, and temporal correlation in control.

One line of work extends the original discretization approach but enriches it with stronger perception and scaling. OpenVLA [96], for example, employs dual vision encoders (SigLIP [104] and DINOv2 [105]) projected into a Llama-2 [106] backbone, enabling action prediction as 256-bin tokens. Despite using just 7B parameters, OpenVLA outperforms larger closed-source counterparts such as RT-2-X [16] (55B), largely due to training on nearly one million robot trajectories from the Open X-Embodiment corpus. SpatialVLA [102] builds on this idea by embedding actions into ''Adaptive Action Grids,'' where motions are discretized into spatially grounded tokens tied to 3D coordinates. This spatially informed design, combined with Ego3D position encodings, improves sim-to-real transfer and allows fine-grained control across environments.

Beyond discretization, a second strand of research embraces diffusion and flow-matching techniques to model the continuous distribution of robot actions. CogACT [98] exemplifies this separation of concerns: it uses a pretrained VLM for perception but delegates trajectory generation to a diffusion-based ''cognition-to-action'' block, capturing multimodal and temporally correlated dynamics. RDT-1B [99] extends this to bimanual manipulation, using a diffusion transformer to predict long action horizons (64 steps) and showing nearly double the performance of models trained without large-scale pretraining. π_0 [103] takes the flow-matching route (as shown in Figure 1), parameterizing an ODE-based transformation on noisy action samples to generate smooth high-frequency control signals (up to 50 Hz). Collectively, these methods address the precision bottlenecks of autoregressive tokenization, enabling dexterity and responsiveness.

A complementary direction introduces cognitive and data-centric innovations. GR00T N1 [101] adopts a dual-system design (shown in Figure 2: a pretrained VLM (System 2, $\sim$1.34B parameters) interprets vision-language input at low frequency ($\sim$10 Hz), while a diffusion-based action transformer (System 1, $\sim$0.9B parameters) outputs motor commands at high frequency ($\sim$120 Hz) using flow matching. Its ''data pyramid'' strategy, integrating internet video, synthetic physics simulations, and real-world robot data, provides robustness to the scarcity and heterogeneity of robot demonstrations. Similar cognition-inspired modularity underlies models like CogACT, which demonstrate substantial improvements over both OpenVLA and RT-2 on standard benchmarks.

In contrast to these works, the GR-2 [100] model adopts a video-first approach. It is pre-trained on a massive volume of internet videos, specifically, 38 million video clips, to predict future frames autoregressively. This process is designed to enable the model to acquire a deep understanding of the world's dynamics, which is then transferred to downstream policy learning during a subsequent fine-tuning stage for action prediction. This methodology treats video prediction as a form of world modeling, providing a strong prior for understanding temporal and physical interactions before learning to control a robot.

Taken together, these advances illustrate a field maturing beyond the early era of discretized action tokens. The emerging consensus is that scalable perception (via VLM backbones and massive datasets) must be coupled with expressive, continuous action modeling (via spatial structures, diffusion, or flow-based architectures). This evolution is transforming VLAs from proof-of-concept systems into dexterous, generalist robot policies with the ability to reason, adapt, and act in complex real-world environments.

The rapid progression of VLA models showcases a significant shift from simple fine-tuning to sophisticated multi-component architectures that deeply integrate visual and linguistic knowledge. These large-scale learning resources are enabling robots to achieve unprecedented dexterity and robust generalization by mastering complex physical and semantic reasoning. Table 4 shows the comparative analysis of the discussed VLA models.

III. APPROACHES TO LEARNING FROM VIDEOS

Researchers have proposed several approaches that adapt videos as the data source for training robots for manipulation tasks. Some of these approaches have borrowed many ideas from computer vision, while a few have also incorporated ideas from language modeling. Substantial advancements have been documented within this domain; nevertheless, a more profound comprehension of the issue and additional exploration into novel learning methodologies, as well as fine-tuning existing ones, are needed to enhance the manipulation skills acquired by robots. The resulting discussions critically assess important literature in this field, highlighting prevailing challenges that impede the capacity of robots to acquire manipulation skills through passive video observation.

The field of video-based robot manipulation encompasses a variety of approaches, each leveraging different techniques to enable robots to learn and perform tasks by observing video demonstrations. This section categorizes these approaches into distinct but interrelated groups, providing a coherent framework for understanding the landscape of methods in this domain. The timeline of the most impactful works and breakthroughs presented in this review is illustrated in Figure 3.

A. FOUNDATIONAL PERCEPTION AND REPRESENTATION METHODS

Effective robot learning from videos relies on establishing perceptual and representational foundations that capture relevant task information and facilitate transfer across

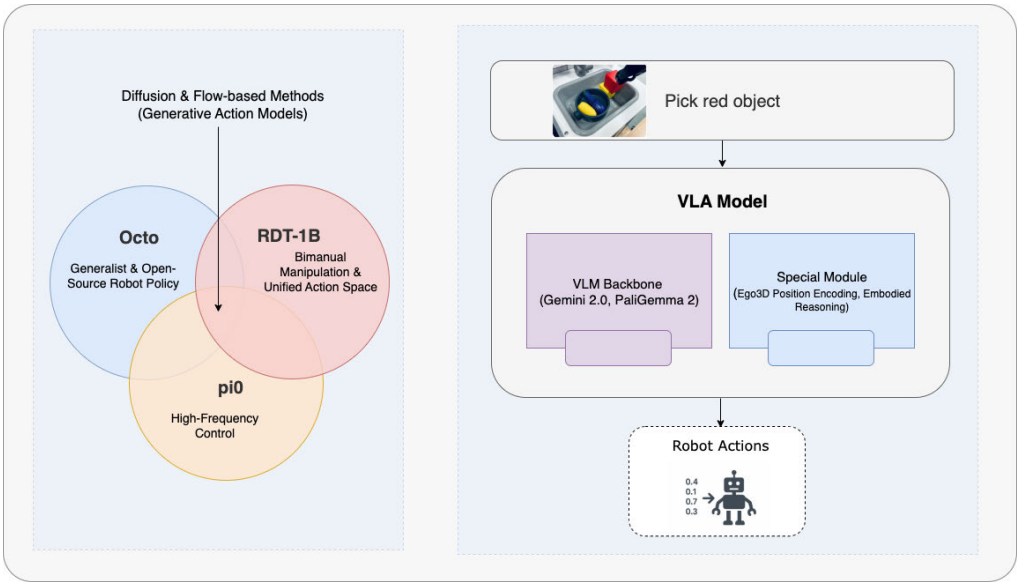

FIGURE 1. Left: Advanced Action Modeling - This group of models uses generative models for action representation, Right: Spatial and Embodied Reasoning - This group of works goes beyond basic visual inputs by incorporating a deeper understanding of 3D space and physical relationships (relevant works: SpatialVLA and Gemini Robotics).

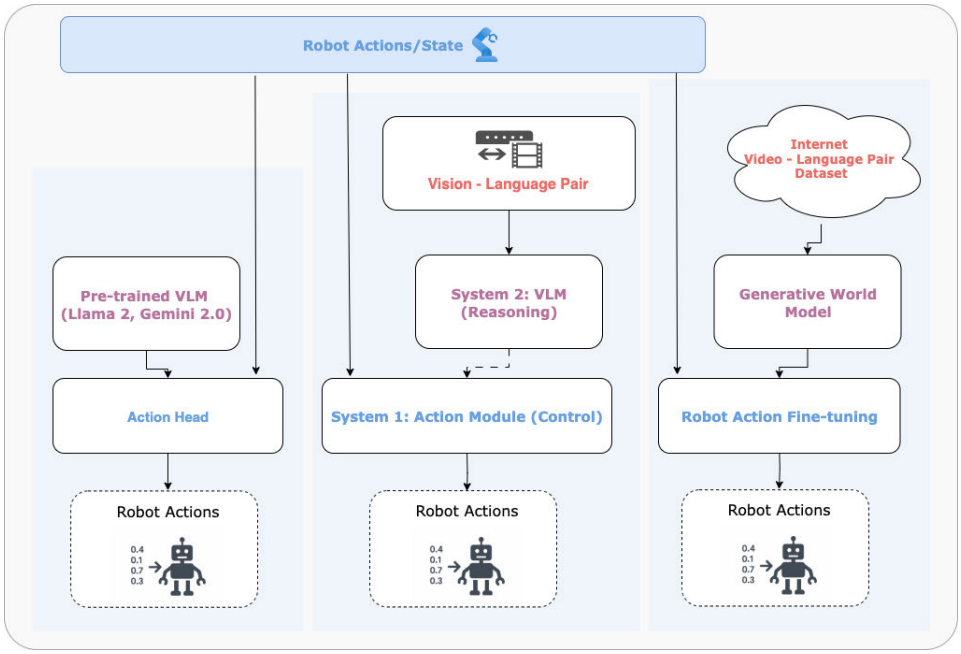

FIGURE 2. This categorizes the models into three groups based on their core design philosophy, showing a high-level view of how each model approaches the problem of VLA modeling. Left: directly adapting a pre-trained VLM for robotic control (relevant works: OpenVLA, Octo, Gemini Robotics, SpatialVLA), Middle: Separates the high-level reasoning from low-level action generation (relevant works: CogACT, GROOT N1, CoT-VLA), Right: Focuses on pre-training on massive datasets of non-robotic videos to learn the underlying dynamics of the world, a form of "embodied physics," before specializing for robot control. (relevant works: GR-2).

domains. On one hand, feature extraction methods focus on deriving meaningful visual or structural representations from raw video, using techniques such as CNN-based encoding or pose and keypoint detection. On the other hand, domain bridging approaches address the gap between human demonstration videos and robotic execution by translating visual and contextual information into robot-compatible representations. Together, these methods provide the essential perceptual groundwork for higher-level learning from videos.

TABLE 4. Comparative overview of state-of-the-art VLA models.

Model	Core VLM/Pre-training	Action Represent. Method	Key Architectural Innovation
OpenVLA [96]	Llama 2 + DINOv2 & SigLIP	Discrete Tokenization	Open-source VLA, efficient fine-tuning (LoRA)
Octo [97]	From scratch (Transformer-first)	Diffusion-based Modeling	Flexible, compositional architecture, open-source
CogACT [98]	Prismatic VLM (Llama 2 + DINOv2 & SigLIP)	Diffusion Action Transformer	Decoupling of "cognition" (VLM) and "action" (DiT)
RDT-1B [99]	From scratch (Diffusion Transformer)	Diffusion-based Modeling	Specialized for bimanual control, Unified Action Space
π_0 [103]	PaliGemma 2	Flow Matching	High-frequency control (50 Hz), multi-expert architecture
GROOT N1 [101]	Gemini 2.0 (distilled)	Diffusion Transformer	Dual-system architecture (System 2 + System 1), data pyramid
SpatialVLA [102]	PaliGemma 2	Adaptive Action Grids	Ego3D Position Encoding, spatial-aware action tokenization
GR-2 [100]	From scratch on internet video	Conditional VAE	Video-generative pre-training, whole-body control (WBC)
CoT-VLA [107]	VILA-U (generative multi-modal)	Hybrid: Causal + Full Attention	Visual chain-of-thought (CoT) reasoning with subgoal images
Gemini Robotics [108]	Gemini 2.0 (VLM)	VLA on Gemini Robotics-ER	Dedicated embodied reasoning model (ER) for perception

1) FEATURE EXTRACTION METHODS

A key step in learning from video is extracting task-relevant features that can guide robot policies. CNN-based approaches learn spatio-temporal representations directly from raw pixels for tasks such as object detection and hand-object interaction. In contrast, pose estimation and keypoint detection methods provide higher-level structural cues by explicitly modeling object geometry and motion dynamics. These two complementary strategies form the basis of video feature extraction for robot learning, as discussed below and shown in Figure 4.

- **CNN-based Feature Extraction**: Early efforts in video-based robot learning primarily relied on Convolutional Neural Networks (CNNs) to extract visual features from demonstration videos, laying the foundation for mapping raw perception to robot action. A representative example is the work of [109], which combined a CNN-based Single Shot MultiBox Detector (SSD) for hand-object interaction detection with a fully convolutional network (FCN) to predict future hand positions. By coupling perception and control, the system directly translated visual cues into motor commands, enabling adaptive robot behavior.

  Subsequent approaches expanded this pipeline with richer perception modules to better capture the complexity of human demonstrations. For instance, [110] augmented CNN-based object detection with OpenPose-based hand localization and greedy video segmentation, allowing the robot to infer collaborative actions and object transfers from relative spatial configurations. Similarly, [111] integrated two-stream CNNs with Mask R-CNN to construct a video parser, which, in combination with a grammar-based execution module, translated visual observations into structured manipulation commands. These approaches illustrate a shift from simple object detection toward frameworks that combine visual parsing with higher-level action reasoning.

  More recent works have pushed CNN-based pipelines toward generalization and scalability. The SWIM framework [112], for example, pretrains on large-scale human interaction videos to learn structured world models of hand-object manipulation. With minimal finetuning on robot data, these representations support goal-conditioned planning, demonstrating the utility of pretraining as a bridge from human demonstrations to robot execution. Previous efforts, such as [113], emphasized compositional representations by pairing CNN-based object detection with grammar-based parsers, highlighting the move toward language-like abstractions of manipulation.

  Addressing domain shift, [114] introduced latent variable models trained on both labeled and unlabeled videos to disentangle shared and domain-specific factors of action. More recently, [115] took this further by eliminating the need for action labels, synthesizing robot-action videos from demonstrations and learning policies directly from raw RGB inputs. These advances mark a trajectory from frame-level feature extraction to structured, generalizable, and annotation-free models of robot manipulation, steadily expanding the robustness and applicability of video-based learning.

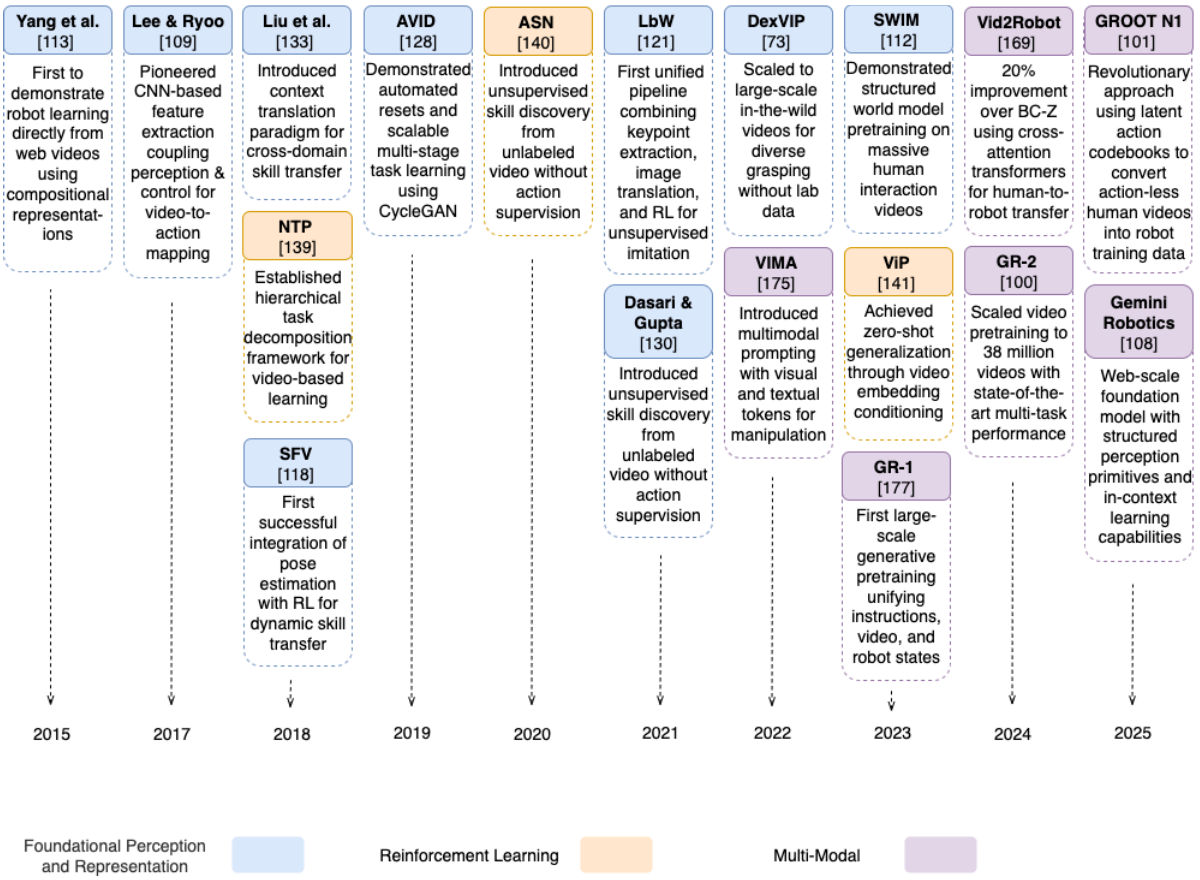

FIGURE 3. A comprehensive timeline organized chronologically by publication year, highlighting the key breakthroughs and milestones each work introduced to the field.

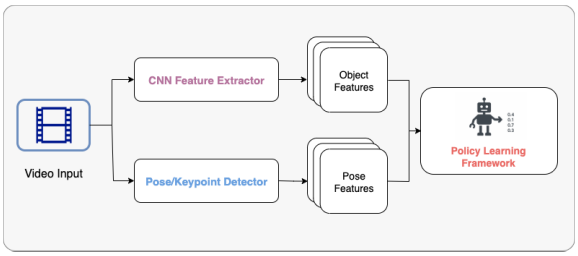

FIGURE 4. Feature extraction methods in video-based robot learning. CNN-based pipelines extract object features and masks, while pose/keypoint methods capture skeletal motion cues, both providing intermediate representations for policy learning.

- **Pose and Keypoint Detection**: While CNN-based pipelines emphasized holistic visual features, a parallel line of work has focused on pose and keypoint detection to capture the fine-grained structure of human motion. Instead of treating video frames as undifferentiated inputs, these methods identify joints, fingertips, or object centers and assemble them into coherent spatial arrangements, providing robots with precise cues about the intent and feasibility of demonstrated actions.

  Several studies leveraged pose estimation to bridge the gap between human demonstrations and robot control directly. For example, [116] extracted detailed hand poses using the 100DOH model and mapped them into robot coordinates with depth sensing, creating strong priors for policy learning. Similarly, [117] proposed a two-stage pipeline that first reconstructs 3D human poses and then retargets them to robot kinematics, ensuring that trajectories remain both functional and feasible. These approaches emphasize embodiment alignment, translating human movements into robot-compatible actions.

  Pose-based reasoning has also proven valuable for dynamic skill transfer. The work of [118] reconstructed actor trajectories from monocular videos and trained reinforcement learning controllers to imitate these motions in simulation, demonstrating how weakly supervised pose estimation can bootstrap skill acquisition. Domain-specific applications, such as [119], have fused Keypoint-RCNN with imitation learning to teach assembly tasks, where reconstructed hand trajectories guided a UR3 robot through precise manipulation.

  Scaling beyond curated datasets, methods like DexVIP [72] exploit large-scale "in-the-wild" human-object interaction videos to learn diverse grasping strategies. By combining human hand pose priors with reinforce-

ment learning, DexVIP achieved generalizable grasping without costly lab-collected data. Meanwhile, lighter-weight approaches combine keypoints with domain adaptation: [120] used EfficientNet-based keypoint detection with CycleGAN translation to align human and robot domains, enabling learning from raw demonstrations. The Learning by Watching (LbW) framework [121] further unified keypoint extraction, image translation, and reinforcement learning into a pipeline that enables unsupervised imitation from human videos without expert annotation.

Together, these works illustrate an evolution from basic pose/keypoint detection toward pipelines that integrate retargeting, domain adaptation, and large-scale pretraining. By capturing structured motion cues, pose-based methods complement CNN-based perception, offering robots interpretable, transferable representations of human actions for imitation and skill learning.

Table 5 shows that feature extraction methods are data-efficient and flexible, well-suited for leveraging large unlabeled video datasets and producing interpretable representations. They work especially well with reinforcement or imitation learning but face challenges with embodiment transfer, scaling to long-horizon reasoning, and domain gaps. Thus, they are valuable for action parsing, object-centric manipulation, and collaboration, though less effective for abstract policy reasoning or seamless transfer.

2) DOMAIN BRIDGING VIA TRANSLATION

When robots learn manipulation skills from human demonstration videos, a significant challenge arises due to the domain shift, primarily stemming from differences in embodiment, visual appearance, and context between humans and robots. Bridging this gap requires intermediate representations and translation techniques that enable robust transfer of skills across domains. Image and context translation methods have emerged as effective solutions, focusing on transforming visual and contextual information to enhance skill transfer, generalization, and adaptability for robotic manipulation.

- **Image Translation**: A major challenge in video-based robot learning is the domain gap between human demonstrations and robot execution. Image translation methods address this by transforming visual data across domains, for instance, converting human-centric demonstrations into robot-perspective images or adapting synthetic renderings into realistic ones. By aligning visual appearances, these methods create a bridge that allows robots to directly interpret and imitate human actions. Classic GAN-based approaches [122], [123], [124] laid the foundation, enhancing the realism of training data and improving robots' ability to generalize manipulation skills [125].

  Within robotics, translation has proven especially powerful for mapping fine-grained object and hand interactions into robot embodiments. For example, [126] introduced a two-module pipeline where a conditional GAN (and domain-invariant networks [5], [35], [36], [114], [127]) predicted visual sub-goals from demonstration videos, while a low-level controller generated corresponding actions. This decoupling of perception and control allowed skill transfer even with unaligned datasets. Similarly, Automated Visual Instruction-following with Demonstrations (AVID) [128] employed CycleGAN [129] to translate human videos directly into robot images at the pixel level, producing instruction images that served as reward signals for reinforcement learning. These works highlighted how direct appearance-level alignment could eliminate the need for manual correspondence between human and robot demonstrations.

  Recent advances have built on these foundations with more expressive architectures. Transformer-based systems, such as the one-shot imitation framework in [130], leveraged self-attention modules to perform unsupervised image-to-image translation, combining goal-conditioned behavioral cloning with deep RL to robustly track and imitate behaviors. Beyond pixel-level mappings, meta-imitation learning methods like A-CycleGAN [131] introduced bi-directional translation between human and robot domains. Coupled with self-adaptive meta-learning, these systems generate imagined robot data to support rapid adaptation with minimal demonstrations, marking a shift toward scalable and flexible cross-domain imitation.
- **Context Translation**: Whereas image translation focuses on aligning visual appearance, context translation tackles the broader challenge of transferring skills across tasks, environments, and viewpoints. This enables robots to adapt behaviors learned in one setting to novel situations with different backgrounds, object positions, or camera perspectives [132].

  One of the earliest examples is [133], which trained a context translation model on paired demonstrations from diverse scenarios. By learning to predict how the same skill looks across contexts, the system enabled a robot to reproduce behaviors in new environments using reinforcement learning. Building on this, [134] developed a context-agnostic task representation paired with a multi-modal inverse dynamics model. By fusing RGB and depth data (and compensating when depth was missing), their system achieved robust action prediction across diverse viewpoints and object configurations.

  Integration with unsupervised video translation further enhanced scalability. The Learning by Watching (LbW) framework [121] combined MUNIT-based [135] video translation with unsupervised keypoint detection, mapping human demonstrations into robot domains without explicit task supervision. Structured keypoint representations extracted from translated videos served as inputs for reinforcement learning, enabling robots to

TABLE 5. Comparison of feature extraction-based approaches for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[109]	Good; real-time collaborative tasks	Moderate; unlabeled videos, needs hand annotations.	Moderate (CNN-based networks)	Direct transfer by visual regression	(+) Unlabeled human videos; (-) Needs initial annotations
[118]	Robust, generalizable dynamic skills (e.g., acrobatics)	Moderate-high; leverages abundant video data	High (deep RL, pose estimation, and simulation)	Pose estimation/motion reconstruction.	(+) Rich video data, dynamic skills; (-) High compute and reconstruction complexity
[114]	Effective tool-use learned purely from observation	Moderate; leverages both observation and interaction data	Moderate-high (latent models and inference)	Learns domain-specific priors	(+) Passive human obs., generalizable; (-) Needs careful latent modeling
[121]	Effective simulation manipulation tasks	Moderate; relies on unsupervised translation and detection	Moderate (images translation, keypoints detection, and RL)	Unsupervised human-to-robot translation	(+) Structured semantics, unsupervised; (-) Simulation only
[72]	Effective dexterous grasping in simulation	High; leverages video priors	Moderate-high (RL, pose extraction)	Hand-pose priors from video	(+) Pose priors, good generalization; (-) Pose extraction limits
[117]	Robust teleoperation & dexterous control in real-time	Low; uses large unlabeled videos	Moderate (pose estimation, neural retargeting)	3D pose-based retargeting	(+) Real-time, low data; (-) Needs reliable pose estimation
[116]	Good; generalizes in many manipulation tasks	High; learns from single human demonstration	Moderate (sampling-based optimization, alignment loss)	Human priors + video alignment	(+) One-shot learning; (-) Sensitive to priors
[112]	Robust for various manipulation tasks	High; few real-world trajectories needed	Moderate-high (world model training and finetuning)	Affordance learning	(+) Few-shot, robust; (-) High model training cost
[115]	Effective performance across manipulation/navigation	High; no action labels needed	Moderate (video synthesis, flow prediction, optimization)	Dense correspondences for action	(+) Inference without action labels; (-) Relies on flow accuracy
[113]	High accuracy on action parsing from unconstrained videos	Moderate; uses large video sets	Moderate (CNN, grammar parsing)	Perception module based	(+) Handles unconstrained data; (-) Needs reliable perception
[119]	Effective in collaborative assembly tasks	Moderate; needs multiple demonstrations	Moderate-high (pose estimation nets, trajectory optimization)	Pose estimation + video retargeting	(+) Accurate pose, task alignment; (-) Video/camera quality sensitive
[111]	Good in multi-object manipulation	High; uses attribute-guided demos	Moderate (CNN, attribute inference)	Attribute-specific retargeting	(+) High object-specific accuracy; (-) Needs attribute extraction
[120]	Effective in simple manipulation tasks	High; 20-30 demos sufficient	Moderate (keypoints extraction, CycleGAN, BC, SQIL)	CycleGAN-based translation	(+) Efficient, robust; (-) Limited task complexity
[110]	Good in collaborative parsing from unconstrained video	Moderate; uses public unlabeled videos	Moderate (YOLO, OpenPose, grammar)	Grammar/symbolic parsing	(+) Generalizable parsing; (-) Action recog. accuracy limits

imitate behaviors under varying contextual constraints. Crucially, both training data for humans and robots were collected via random demonstrations rather than expert labels, lowering the barrier for large-scale data collection.

Image and context translation offer complementary strategies to bridge the human-robot domain gap. Image translation aligns the robot's visual perception with demonstrations, while context translation adapts actions across environments, viewpoints, and tasks. Progress has evolved

TABLE 6. Comparison of image and context translation-based approaches for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[133]	Good; high sim task success; outperforms GAIL/TPIL	Moderate; needs multiple human video demos from varied contexts and 100k+ samples during RL	Moderate; translation model + visual encoder + RL (TRPO or GPS)	Weak; assumes human and robot use same tools and similar viewpoint to reduce domain gap	(+) Learns from raw video; (-) Requires morphology/demo alignment
[126]	High; 60-75% real robot success; outperforms end-to-end and DAML baselines	High; reusable low-level controller; fewer task-specific samples needed	Moderate; U-Net + ResNet-based goal generator + inverse model; trained separately	Moderate; uses GAN-based sub-goal transfer	(+) Hierarchical; modular/sample-efficient; (-) Goal generator task-specific
[128]	High; 80-100% multi-stage task success	Very high; learns full tasks with ~20 mins of human video and 180 mins of robot practice	High; CycleGAN + structured latent model + MPC + classifier-based reward	Strong; translates entire human demo to robot via CycleGAN without paired data	(+) Automated resets, scalable; (-) CycleGAN requires good translation data, per-task model
[130]	High; 88.8% pick-place success	High; 3x less data needed	High; Transformer with inverse dynamics and keypoint loss	Strong; self-supervised sim-to-real	(+) Excellent transfer, modular; (-) High data/compute
[121]	Strong; comparable or superior to AVID/Classifier-based methods	High; learns from single human demo per task	Moderate; CycleGAN + keypoint extraction + RL	Strong; keypoints structure for transfer	(+) Avoids artifacts, efficient; (-) Needs robust translation model
[131]	Strong; matches DAML without robot demos	High; only human video needed during training	High; A-CycleGAN + meta-learning	Strong; A-CycleGAN handles shifts	(+) No robot demos; (-) Relies on good latent/action inference
[134]	Good; beats baselines on stacking in different contexts	Moderate; uses paired data and depth prediction	Moderate-high; 4-model pipeline with depth estimation	Moderate; via context translation and multimodal input	(+) Strong cross-context performance with RGB-D; (-) Needs depth prediction model and alignment

from early GAN-based pixel translation to transformer and meta-learning frameworks, moving toward scalable, annotation-efficient, and robust pipelines for unstructured settings.

Table 6 illustrates how these methods address visual and semantic gaps, enabling sim-to-real and cross-domain transfer from raw or unpaired video data. However, challenges remain in precise action alignment, artifact reduction, and managing embodiment mismatches. Despite these hurdles, context translation remains critical for advancing generalization and real-world adaptability.

B. REINFORCEMENT LEARNING (RL) APPROACHES

Reinforcement Learning (RL) provides a powerful paradigm for enabling robots to acquire manipulation skills through interaction, optimizing behavior by trial and error guided by reward signals. In the context of learning from video, recent research has explored how RL can be combined with rich perceptual inputs extracted from demonstrations. This integration allows robots to benefit both from their exploratory learning and from structured guidance provided by human expertise in video data. A depiction of the RL-based subcategories is shown in Figure 5.

1) VISUAL RL WITH FEATURE EXTRACTION

Early approaches sought to ground RL in visual representations derived from video, leveraging feature extraction to transform demonstrations into useful training signals. For example, the video parsing framework of [136] combined Mask R-CNN with a dedicated hand-object detector to build coarse 3D scene representations from human demonstrations. These representations were aligned across trajectories and used to generate dense reward signals, guiding RL policies toward precise motor execution. Similarly, [137] emphasized extracting tool motion from instructional videos, aligning simulated environments with human demonstrations, and employing trajectory optimization to bridge from visual guidance to executable robot policies. In both cases, parsing video into trajectories provided structured signals that made RL training more efficient.

A persistent challenge, however, is that raw video demonstrations typically lack explicit action or reward labels, and domain shift can make learned policies brittle. To address this, [138] introduced a hybrid approach that combined offline observational data with online interaction. Their system maintained dual replay buffers, one for action-free video observations and another for action-conditioned robot experience, and learned inverse models over compressed,

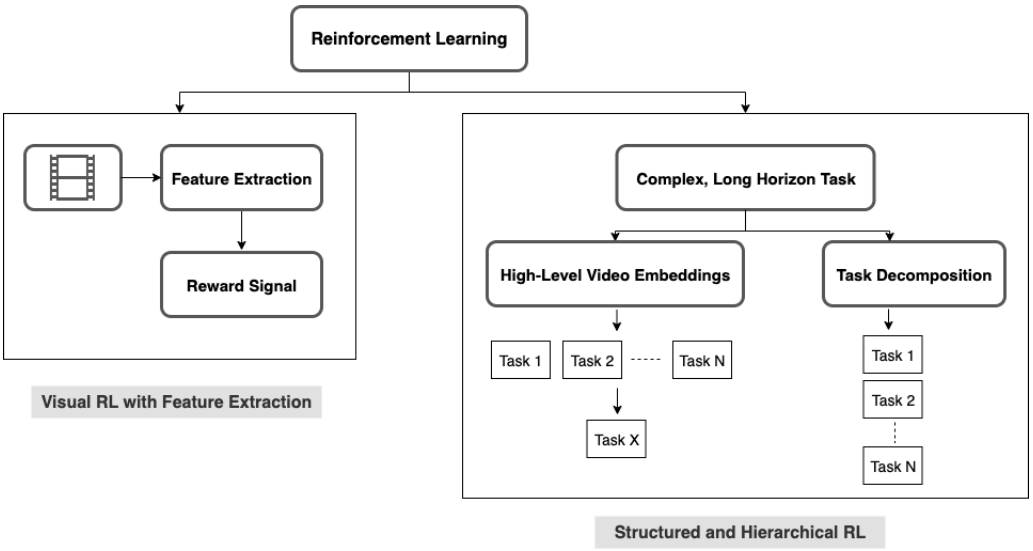

FIGURE 5. Reinforcement learning paradigms in video-based robot learning - Left: visual RL with feature extraction, grounding policies in parsed video features; Right: structured and hierarchical RL, learning high-level video embeddings for multiple tasks, and decomposing long-horizon tasks into subtasks and primitive skills.

domain-invariant features. This design allowed robots to infer missing actions and transfer knowledge across domains, advancing RL toward robustness in heterogeneous video settings.

2) STRUCTURED AND HIERARCHICAL RL

As research progressed, focus shifted from single-task policies toward generalizable and scalable frameworks. Neural Task Programming (NTP) [139] exemplified this trend by decomposing tasks into hierarchical compositions of primitive actions, parameterized by trajectories or video demonstrations. Through meta-learning strategies, NTP enabled robots to adapt policies quickly across diverse manipulation tasks. Likewise, Adversarial Skill Networks (ASN) [140] learned a task-agnostic skill embedding space from unlabeled, multi-view observations. By leveraging adversarial and metric learning objectives, ASN eliminated the need for explicit action supervision, highlighting how representation learning can facilitate transferable skills.

More recently, Video-conditioned Policy Learning (ViP) [141] has advanced the integration of RL with large-scale video data. ViP conditions policies directly on video embeddings of demonstrations, using a supervised contrastive encoder trained on human activity datasets like Something-Something-v2 [83]. At inference, ViP retrieves task embeddings via nearest-neighbor search and conditions policy learning on them, enabling multi-task and zero-shot generalization. By directly grounding RL in high-level video representations, ViP points toward scalable solutions where robots can learn broad repertoires of behaviors from human video libraries without paired training data.

Table 7 highlights the strength of reinforcement learning in tackling complex, long-horizon tasks and achieving robust sim-to-real transfer, particularly in hierarchical and multi-task manipulation. However, these gains come with high computational and data costs, along with sensitivity to reward design and training stability, limiting RL's practicality despite its adaptability.

C. IMITATION LEARNING (IL) APPROACHES

Imitation Learning (IL) is one of the most prominent strategies for training robot manipulation policies. In contrast to reinforcement learning, which depends on trial-and-error exploration, IL enables robots to acquire skills directly from human demonstrations. This reduces the need for carefully engineered reward functions and often yields more sample-efficient, generalizable policies, particularly valuable in data-limited or real-world settings. Figure 6 shows the different taxonomy of IL based approaches, and they are discussed in details below.

1) BEHAVIORAL CLONING (BC) AND VARIANTS

A large body of IL research builds on behavioral cloning (BC), where visual inputs from demonstration videos are mapped directly to robot actions. Modern approaches typically employ CNN-based encoders for feature extraction, combined with deep policy networks for action prediction.

Recent extensions of BC have introduced more flexible task representations. For example, BC-Z [10] supports both video- and language-specified tasks, enabling zero-shot generalization. Its architecture combines a ResNet18 [142] backbone with FiLM layers [93] for task conditioning, and incorporates human-in-the-loop corrections via teleoperation.

Other works have focused on data efficiency. TecNets [2], [143] encode demonstrations into compact embeddings that

TABLE 7. Comparison of reinforcement learning-based approaches for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[139]	High; excels in hierarchical tasks (stacking, sorting)	High; few demos needed	Moderate (due to hierarchical decomposition and recursive calls)	Good; task decomposition, program induction	(+) Hierarchical generalization, modular, few-shot; (-) Sensitive to low-level API/collision
[140]	High; learns complex tasks from video	Moderate; uses unlabeled video data	Moderate (metric and adversarial learning)	Good; adversarial skill-transfer embeddings	(+) Transferable embeddings, unsupervised skill disc.; (-) Needs careful metric learning
[138]	High; strong on vision-based tasks	High; leverages human videos, less robot data	Moderate (inverse model training, adversarial domain confusion)	Excellent; domain-invariant embedding	(+) Good domain generalization, sim-to-real; (-) Dependent on inverse model accuracy
[136]	Good; effective for object manipulation tasks	Moderate; few demo videos needed	Moderate (differentiable rendering, RL training)	Excellent; 3D state estimation transfer	(+) Robust 3D generalization; (-) Approximation errors for complex scenes
[137]	High; effective for tool manipulation (spade, hammer, scythe), 100% success rate	High; requires only single video demonstration	Moderate (trajectory optimization + PPO, alignment sampling up to 20k iterations)	Excellent; morphology-agnostic via tool-centric approach	(+) Tool-focused transfer, works across robot morphologies, real robot validation; (-) Requires sparse reward environment, limited to stick-like tools, needs good video visibility
[141]	Excellent; zero-shot manipulation from human videos	Highly efficient; leverages large human datasets	Moderate (pretrained video embedding, efficient inference)	Excellent; pretrained action embeddings	(+) Zero-shot generalization, efficient inference; (-) Relies on pretrained embeddings

condition policies for rapid adaptation, while Multiple Interactions Made Easy (MIME) [144] scales imitation through a large demonstration dataset, pairing VGG-based [145] visual encoders with LSTM-based [146] trajectory prediction.

Beyond direct cloning, IL can also be framed as an inverse reinforcement learning (IRL) problem, where the goal is to infer a cost function from demonstrations. A key direction here is Imitation from Observation (IfO), which avoids reliance on expert action labels. Generative Adversarial Imitation from Observation (GAIfO) [147] exemplifies this approach by recovering expert-like policies from observed state transitions alone, providing a more scalable alternative to traditional IRL.

2) META-IMITATION AND FEW-SHOT IL

To improve generalization, recent research integrates IL with meta-learning and few-shot learning. The central idea is to train on diverse tasks so that robots can quickly adapt to novel ones from only a handful of demonstrations [127], [148].

Zero-shot imitation has been explored by [149], who combined exploration-driven policies with forward-consistency losses to enable imitation without labeled demonstrations. Similarly, MOSAIC [150] employs multi-task architectures with self-attention and temporal contrastive modules, enabling robust representation learning and improved task disambiguation. Meta-learning methods such as Model-Agnostic Meta-Learning (MAML) have also been adapted for IL. Reference [5] trained policies across varied prior tasks so that only a single new demonstration is required for adaptation at inference. This one-shot generalization illustrates the promise of meta-imitation in tackling domain shift and data scarcity.

3) CROSS-DOMAIN AND HUMAN-TO-ROBOT TRANSFER

A parallel line of work addresses cross-domain transfer, particularly from human video demonstrations to robot execution. The work done in [151] demonstrated multi-step task learning by localizing human-demonstrated actions within supplemental videos, combining BC, IRL, and RL for accelerated refinement. The authors in [152] proposed a more direct strategy, applying BC on raw human videos without explicit domain adaptation. By leveraging end-to-end training with Adam optimization [153] and occlusion-based perspectives (e.g., eye-in-hand views), their method mitigates domain shift and enables direct policy transfer. Other approaches incorporate auxiliary supervision signals.

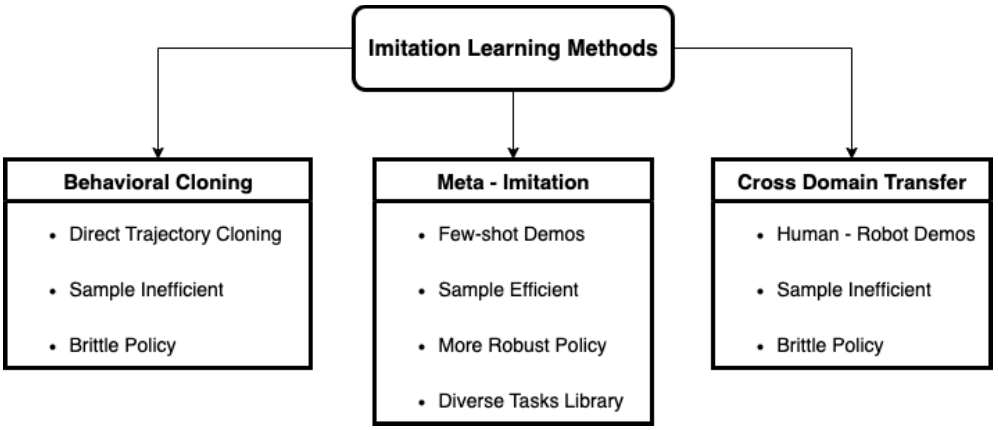

FIGURE 6. Taxonomy of imitation learning approaches - Left: behavioral cloning from raw videos, (Middle): meta-imitation from few-shot demonstrations, and Right: cross-domain transfer from human-to-robot videos.

The Watch, try, learn (WTL) framework [154] integrates visual meta-learning with binary success/failure feedback, enabling policy refinement in sparse-reward settings. Graph-based representations such as Visual Entity Graphs (VEGs) [155] further enhance transfer by explicitly modeling spatial and temporal relations between objects and actions, enabling single-demonstration learning in real robotic systems.

Imitation learning approaches, as summarized in Table 8, stand out for their sample efficiency and direct use of expert demonstrations, making them ideal for scenarios where rapid skill acquisition is critical. The comparative analysis shows that advances in meta and one-shot learning have further enhanced their ability to generalize from limited data. However, the table also highlights persistent issues with embodiment alignment and generalization, especially across domains or when transferring from human to robot. Many methods remain primarily evaluated in simulation. In practice, imitation learning excels for tasks with readily available demonstration data but may require augmentation or hybridization for broader applicability.

D. HYBRID APPROACHES

Hybrid models have emerged as a powerful paradigm in robot learning, combining RL, IL, and complementary techniques to overcome the limitations of relying on a single method. Early work demonstrated the effectiveness of augmenting IL with RL-based fine-tuning, such as attention-driven imitation guiding RL policies [156] and pose-driven motion imitation refined via RL [118]. RL-based residual corrections to demonstration trajectories [157] has also stood out as an approach to adapt video demonstrations to robot trajectories. Other efforts leverage cross-domain priors [158], unaligned video demonstrations [159], or adversarial third-person imitation [160] to improve generalization and robustness. More recent approaches combine IL, RL, and control-theoretic methods such as model predictive control (MPC) for more efficient skill transfer [161]. By integrating multiple learning strategies, these approaches enhance data efficiency, adaptability, and robustness, particularly in complex manipulation tasks.

1) HYBRID RL AND IL

A central motivation for hybrid RL-IL approaches is to combine the sample efficiency of imitation with the exploration and optimization strengths of RL. Attentive Task-Net [156] exemplifies this direction by integrating a self-supervised attention network for viewpoint-invariant imitation with an RL policy optimized using Deep Deterministic Policy Gradient (DDPG) [162]. A CNN-based embedding network learns task-relevant visual features, which are refined by spatial attention and used to guide RL agents, producing policies that balance efficient imitation with long-horizon optimization. Similarly, Skills From Videos (SFV) [118] extracts human poses from monocular video, reconstructs reference motion trajectories, and then trains deep RL policies to follow these trajectories in simulation, demonstrating how perception-driven imitation can be combined with RL to achieve robust skill transfer across embodiments and environments.

2) DOMAIN ADAPTATION HYBRIDS

Hybridization has also proven critical for addressing domain adaptation. Semantic Transfer Accelerated RL (STAR) [158] leverages demonstrations from different domains by encoding action sequences with a conditional VAE, pretraining low-level policies, and exploiting semantic priors to align states across tasks. By minimizing the KL divergence between semantically similar states and exploiting temporal context, STAR achieves efficient skill transfer across domains.

Another influential line of work uses unaligned video demonstrations. Reference [159], for example, trained agents to play Atari by leveraging YouTube videos without direct domain alignment. Through self-supervised temporal distance classification and representation learning, their method enabled human-level performance in challenging exploration tasks, demonstrating that hybrid IL-RL models can generalize even from noisy, cross-domain data.

TABLE 8. Comparison of imitation learning-based approaches for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[147]	High, matches IL from state-only, no actions needed	Low; needs large robot data; not suitable for human demonstration or low-data settings	Moderate (model-free RL training with discriminators)	None; needs shared state/action space	(+) No action labels, partial demos; (-) No human-to-robot transfer, no raw vision use
[144]	Not applicable; introduces a large-scale multi-task dataset	High; supports low-data learning algorithms	N/A (dataset paper)	Human/robot paired demos	(+) Largest video-trajectory dataset; (-) Limited real-world variety, kinesthetic only
[155]	Good; achieves robust performance using graph-structured perception from a single human video demonstration	High; one-shot imitation, no robot data required at test time	High (graph generation, matching over time, entity detectors)	Strong; dynamic graph aligns hand/object keypoints	(+) One-shot, no instrumentation; (-) Needs reliable keypoint detection, compute heavy
[143]	High; effective at one-shot imitation in simulation using image-based embeddings	Moderate; needs diverse pairs for training; efficient at test time due to embedding reuse	Moderate (encoder + BC + embedding loss)	Partial; generalization by visual embedding	(+) General/one-shot; (-) Sensitive to domain gap, real-world untested
[150]	High, strong across 7 tasks, 61 using a unified transformer-based policy	Moderate; shared transformer, task data needed	High (transformer, temporal loss)	Partial; generalizes robot arms, not human-to-robot	(+) Unified policy, generalizes; (-) Complex pipeline, demo diversity required
[10]	Good; 44% zero-shot success on 24 unseen real-world tasks using video/language input	Moderate; requires thousands of demos, few/zero-shot efficient after pretraining	Moderate (ResNet encoders + FiLM conditioning + BC)	Strong; video/language embedding bridges modalities	(+) Flexible input, strong generalization; (-) Needs quality pretrained embeddings
[154]	Moderate, learns reward from online human video for RL	Low; video + environment interaction for reward inference and training	High (reward learning, video encoding, RL)	Weak; learning sensitive to visual mismatch	(+) Autonomous reward from video; (-) Artifacts/noise hurt policy
[5]	High; strong one-shot performance on sim and real-world robotic tasks using gradient-based adaptation	High; only one test-time demonstration required after meta-training on human+robot data	High (MAML optimization, visual encoder, adaptation at inference)	Moderate; domain-adaptive feature aligns human-robot	(+) Fast adaptation, flexible input; (-) Needs meta-training, paired modalities
[152]	High; 58% improvement in real-world manipulation	High; avoids expert robot demos by using inverse model trained on play data to label human videos	Moderate (inverse, BC, masking)	Strong; masking + camera perspective closes gap	(+) Robust/scalable, no robot demo; (-) Masking can omit context
[149]	High; real and simulation tasks via goal images, no expert actions	High; no expert demos, needs unsupervised pretraining	Moderate-high (exploration, forward consistency)	Moderate; learns goal-conditioned skills during exploration	(+) Unsupervised, strong generalization; (-) Relies on exploration/visual similarity
[151]	Moderate-High; achieves multi-step task execution from one segmented demo + auxiliary videos	High; few-shot learning via video snippets/meta-learned localization	Moderate; relies on few-shot video classification (MAML/Reptile) and PPO for policy training	Moderate; generalizes across unseen colors and robot arms, but not evaluated on human-to-robot transfer	(+) Multi-step one-shot, robust reward inference; (-) No real-robot transfer, sim only

Residual learning further illustrates the utility of hybrid models, and works such as [157] leveraged it to train a residual RL agent to refine noisy human hand poses for dexterous manipulation, with adversarial imitation ensuring corrections remain physically plausible. Similarly, [160] introduced third-person imitation learning, where RL combined with a GAN-based cost function recovery enabled imitation from videos with differing viewpoints and embodiments. More

recent work [161] extends this paradigm by leveraging task family priors and temporal abstractions extracted from demonstrations, alongside sampling-based Model Predictive Control (MPC) for safe trajectory generation.

Overall, hybrid approaches advance robot learning by combining IL's efficiency, RL's adaptability, and modern perception-control flexibility. This synergy enables robots to learn robust skills from a few demonstrations, transfer across domains and embodiments, and adapt policies online—bringing real-world deployment closer.

Table 9 shows that hybrid models, integrating RL and IL, handle noisy demonstrations, sparse rewards, and domain adaptation well, excelling in cross-domain and viewpoint-variant tasks. However, they introduce design complexity and lack broad real-world validation. Overall, hybrid methods strike a balance between RL and IL strengths but remain best suited for simulation or moderately complex tasks.

E. MULTI-MODAL APPROACHES

Robotic manipulation in unstructured environments requires the ability to interpret complex sensory signals and ground them in actionable policies. Relying on vision alone is often insufficient, as tasks typically involve abstract goals, contextual reasoning, or subtle cues that exceed purely visual perception. To address this, researchers have increasingly turned to multi-modal approaches (as shown in Figure 7, where vision is combined with other input streams such as touch, proprioception, and, notably, natural language. By leveraging these complementary modalities, robots gain a richer and more holistic understanding of manipulation tasks, enabling them to follow nuanced instructions, reason about interactions, and generalize to unseen scenarios.

1) VISION-LANGUAGE GROUNDING

The earliest wave of vision-language methods focused on establishing a direct connection between natural language instructions and robotic action. For instance, [163] fused natural language instructions with scene images to create task-specific embeddings that guided a policy network in generating robot trajectories. To enrich semantic understanding, the model leveraged a video-based action classifier trained on the Something-Something dataset [83], aligning robot behavior with human demonstrations.

Before that, some researchers approached the topic from the standpoint of commonsense reasoning in robotics. Reference [164] combined attention-based VLMs with ontology systems to represent manipulation concepts in time-independent semantic structures. Their introduction of the Robot Semantics Dataset and spatial attention mechanisms for action captioning laid the foundation for knowledge-graph reasoning in manipulation.

There has also been progress with methods that linked low-level perception to higher-level linguistic descriptions. For example, [165] proposed a dual-model architecture: a grasp detection network (GNet) that computed object grasps and a captioning network (CNet) that translated demonstration videos into commands. Here, language was not only a means of communication but also a scaffold for structuring perception and action.

Building on these ideas, works such as Watch and Act [166] introduced pipelines where video captioning and robot planning were tightly coupled. Demonstration videos were first converted into textual instructions, which were then grounded in visual perception modules (e.g., segmentation) and executed through RL-based controllers. This marked a shift toward systems capable of seamless transitions from naturalistic demonstrations to executable actions.

Generalization has also been an active focus in vision-language learning. Reference [167], for instance, enabled robots to perform zero-shot imitation of both single-agent and collaborative tasks from YouTube videos. Instead of generating commands, their framework constructed action grammars and action trees, providing a structured yet flexible approach to representing novel behaviors. These advances highlight the progression from basic grounding of instructions to more scalable frameworks that support open-ended learning.

2) VISION-ACTION ALIGNMENT VIA MULTI-MODAL REPRESENTATIONS

While grounding language in perception was a critical first step, recent advances extend beyond mapping words to actions toward building unified multimodal representations. These approaches jointly embed videos, states, and textual instructions into shared spaces for reasoning and control [15], [16]. This integration moves the field from task-specific pipelines to generalizable frameworks that bridge perception and action more seamlessly.

Early attempts at such alignment explored direct translation of demonstrations into robot instructions. Reference [168] employed CNNs and RNNs to convert visual features into grammar-free command descriptions, demonstrating how semantic comprehension could augment traditional imitation learning. Subsequent approaches like Vid2Robot [169] refined this idea by using cross-attention to align video prompts with a robot's state, with contrastive losses ensuring robust representation learning. These works underscored the value of multi-modal alignment in supporting long-horizon reasoning and motion transfer.

The ability to predict future dynamics further broadened the scope of these models. Reference [170] proposed predicting "general flow" 3D trajectories of object points from RGB-D video and language input, enabling skill transfer across embodiments and morphologies. Similarly, [130] leveraged transformers for one-shot imitation from videos, introducing an inverse dynamics loss to stabilize self-attention and improve policy adaptation. Together, these models illustrate a progression from mapping demonstrations to generalizable trajectory prediction.

TABLE 9. Comparison of hybrid approaches for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[160]	Succeeds on pointmass, reacher, and inverted pendulum (via 3rd-person demos) tasks	Efficient: no action/state alignment needed	Moderate: adversarial training + domain confusion	Explicit; learns domain-agnostic features	(+) Unsupervised, no action labels, handles domain gap; (-) Simple tasks only
[159]	Achieves and surpasses human-level in difficult Atari games (e.g., Montezuma) from video	Very efficient: single video demo is sufficient	High: deep self-supervised embeddings + RL training	Strong; domain-invariant video embeddings	(+) Solves sparse-reward/complex tasks, robust to visual gap; (-) Heavy compute, complex train
[118]	Learns high-fidelity dynamic skills (locomotion, acrobatics)	Efficient: uses public video data, minimal motion capture	High: deep pose estimation + RL with adaptive curriculum	Robust; physics-based policy handles noise	(+) Learns from unstructured video, retargets skills; (-) Needs good pose est., sim only
[157]	Improves success in dexterous VR manipulation and in-the-wild hand tracking	Needs mocap data for initial training, but less than full demo collection	Moderate (model-free hybrid RL + IL, hand pose estimation)	Residual policy corrects pose errors	(+) Physics-based, robust to estimation errors; (-) Needs mocap dataset for train
[158]	Matches demo-accelerated RL for long-horizon kitchen tasks	Very efficient: <3 minutes human video enables long-horizon skill transfer	Moderate (semantic skill extraction, RL)	Robust; semantic imitation, cross-domain	(+) No in-domain demos, scalable, generalizes; (-) Needs offline skill extraction
[161]	High; enables one-shot fabric manipulation from video	Highly efficient: single demo + sim prior needed	Moderate: sim pretraining + MPC for policy optimization	No strict sim-to-real; scene-level alignment	(+) No risky real-world explore, efficient; (-) Focused on fabric, needs scene prior
[156]	Outperforms SOTA in pouring task imitation (lower error, fewer iterations)	Sample-efficient due to self-supervised and attention-guided feature learning	Moderate (CNN encoder + attention module + RL)	Attention-guided: view/background invariant	(+) Robust to clutter, learns focused features; (-) Needs multi-view, task-tuning

With internet-scale video data becoming available, generative architectures have pushed multi-modal learning into broader domains. Reference [171] trained an image transformer with a conditional variational autoencoder (C-VAE) to anticipate human actions and object interactions from diverse online videos. The resulting model achieved zero-shot transfer to novel lab settings, demonstrating the potential of pretraining on open-world data. Building on this principle, PLanning-EXecution (PLEX) [172] introduced a planner-executor transformer framework that separates high-level activity sequencing from low-level action execution, supporting multi-task generalization even in low-data regimes.

Taken together, these advances chart a clear trajectory: from early instruction-to-action systems to transformer-based architectures unifying vision, language, and action, and finally to large-scale pretraining for robust generalization. Multi-modal representations have become a cornerstone of next-generation robotic intelligence, enabling grounding, abstraction, and synthesis for versatile real-world manipulation.

Table 10 highlights the strong ability of these methods to integrate vision, language, and sometimes other modalities, supporting flexible, language-driven, and zero-shot manipulation. They excel at generalization, especially for long-horizon and open-vocabulary tasks, but face challenges in model complexity, data demands, and training cost. Thus, while powerful and general, multi-modal approaches remain constrained by the need for rich datasets and significant computational resources.

3) MULTI-MODAL TRANSFORMERS AND LARGE-SCALE FOUNDATION MODELS

While early multi-modal approaches established the feasibility of combining perception, action, and language, their scalability remained limited. The arrival of large-scale transformer architectures and language-conditioned policy models has since transformed the landscape, providing the representational capacity and flexibility needed to tackle long-horizon, multi-task robot learning. These advances are driven by a common intuition: that transformer-based models,

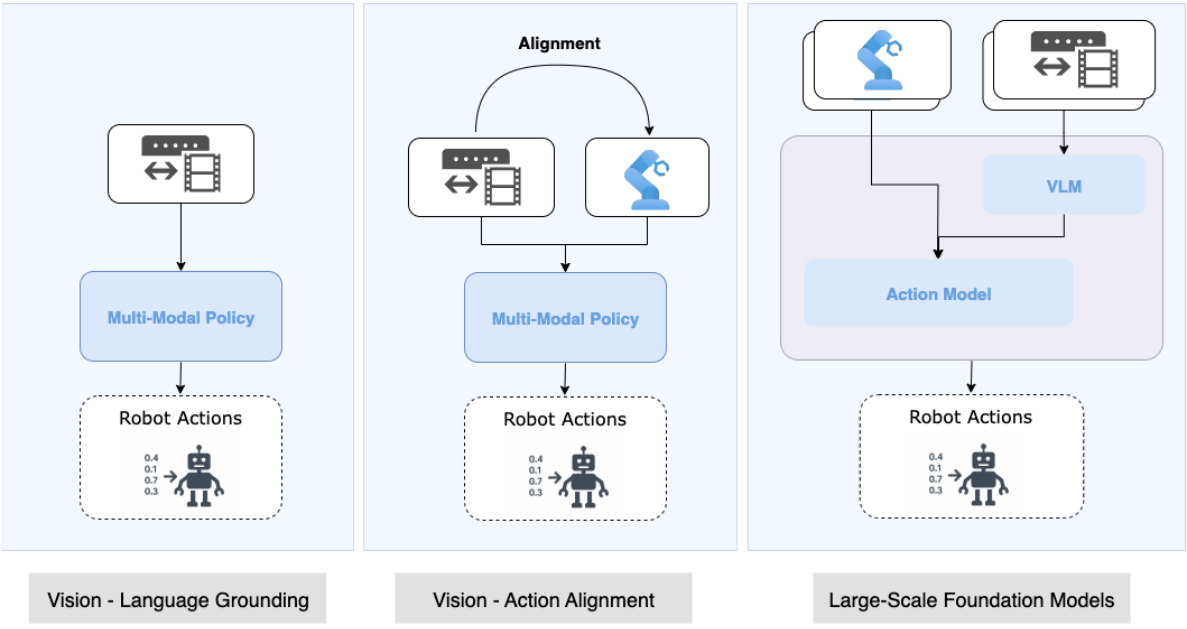

FIGURE 7. Evolution of multi-modal architectures in robot learning - Left: vision-language grounding, (Middle): vision-action alignment with shared embeddings, and Right: large-scale transformer-based foundation models (VLAs).

pretrained on diverse multimodal data, can capture the temporal, semantic, and structural regularities necessary for robust manipulation policies.

Initial breakthroughs demonstrated the power of hierarchical attention mechanisms for grounding multimodal signals. Hierarchical Universal Language Conditioned Policies (HULC) [173], for example, employed a hierarchical transformer encoder with contrastive alignment of video and language embeddings to support long-horizon manipulation. Evaluated on Composing Actions from Language and Vision (CALVIN) [174], HULC achieved strong generalization across tasks, marking one of the first demonstrations that multimodal transformers could scale beyond narrow, single-task pipelines. Building on this foundation, models such as VisuoMotor Attention (VIMA) [175] introduced transformer-based architectures that process multimodal prompts composed of visual and textual tokens. By incorporating pretrained language models into an encoder-decoder system, VIMA was able to perform data-efficient policy learning across a wide range of manipulation tasks, showing how prompting can unify task specification and execution. Similarly, DigKnow [176] leveraged LLMs to extract layered knowledge from scene graphs in human demonstration videos, enabling retrieval and correction mechanisms that improve generalization to novel task instances.

A major shift occurred with the adoption of generative pretraining paradigms. GR-1 [177], for instance, unified instructions, video observations, and robot states into a single predictive architecture. By pretraining on massive video datasets and fine-tuning on robot data, GR-1 achieved state-of-the-art results on challenging manipulation benchmarks. Similarly, Video-based Policy learning framework via Discrete Diffusion (VPDD) [178] demonstrated how discrete diffusion models could compress video data into latent tokens, predict future video dynamics, and then fine-tune with limited robot-labeled data. These approaches established video pretraining as a key enabler of scalable robot learning.

Recent VLA models have pushed these ideas further by leveraging hundreds of millions of internet videos to endow robots with broad physical priors and temporal reasoning. For example, GR-2 [100] was pretrained on 38 million videos to learn conditional temporal prediction, then fine-tuned on robot demonstrations to achieve state-of-the-art multi-task performance (as shown in Figure 2). GROOT N1 [101] extended this idea with a latent action codebook, enabling even action-less human videos to be repurposed as robot data. By generating synthetic "neural trajectories" through video generation models, it amplified its training corpus dramatically and, with a diffusion transformer, achieved robust multimodal control across diverse robots. Other models have explored new mechanisms for integrating human video data into robot learning. CoT-VLA [107] introduced a visual chain-of-thought framework, predicting subgoal images before producing corresponding actions. This allowed training on large-scale human activity datasets such as Epic-Kitchens, despite the lack of action annotations, improving long-horizon reasoning. Gemini Robotics [108], meanwhile, extended Gemini 2.0 with an Embodied Reasoning (ER) layer. By training on massive web video corpora, it acquired structured perception primitives, such as 3D object detection, grasp prediction, and trajectory reasoning, that could be adapted to novel tasks with only in-context prompting or minimal fine-tuning.

TABLE 10. Comparative analysis of vision-language grounding and vision-action alignment approaches for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[168]	Strong hierarchical task generalization and adaptation	Efficient due to hierarchical reuse	Moderate (modular network execution)	Partial handling via sub-program abstraction	(+) Good Compositionality and generality; (-) Needs task sketch annotations
[165]	Good for long-horizon tasks	Needs many demos	High due to GCNs and temporal reasoning	Not directly addressed	(+) Task-structure aware; (-) High memory/sample cost
[167]	Good zero-shot on unseen tasks due to pretraining with human video)	Very efficient; human video only for most of training	Moderate-high	Object-centric/temporal cues bridge gap	(+) Pure visual imitation; (-) Needs temporal coherence
[164]	Good for vision-language manipulation	Relatively data-efficient due to CLIP feature reuse	Moderate-high	Indirect via pretrained Vision-Language features	(+) Language generalizes; (-) Fails on vague instructions
[169]	High; 20% improvement over BC-Z on real robots tasks	Data-efficient; due to paired video/trajectory data	High (large transformer with cross-attention over videos)	Contrastive loss/data pairing bridge gap	(+) Imitate from human demos; (-) Data collection bottleneck
[163]	High; learns 78 tasks from human videos; strong simulated generalization across instructions/scenes	Efficient: self-supervised RL, no teleop demos	Moderate (RL + supervised imitation; video classifier training)	Video reward model bridges sim2real	(+) Diverse skills from unstructured video; (-) No real-robot validation, open-loop
[166]	High; validated on 24 objects, 8 actions, and real robot tasks	Moderate: relies on synthetic data for segmentation; video demos for policy learning	Moderate (captioning, segmentation, RL-based affordance)	Captioning/modular policy reduce gap	(+) No expert action data, real-world diversity; (-) Pipeline complexity, caption accuracy dep.
[170]	81% success across 18 real-world tasks	No robot data; trained on human videos only	Low (simple policy; training cost in flow model)	Very low gap via flow affordance	(+) Embodiment-agnostic; (-) Less expressive than trajectories
[130]	High; 2x better than prior pick-place baselines in sim (16 tasks)	Moderate; needs 1600 demo-context pairs across tasks; one-shot at test time	Moderate; Transformer-based encoder-decoder with ResNet and multi-loss training (BC + inverse + keypoint loss)	Partial; visual transfer across Sawyer/Panda in sim	(+) Strong generalization via attention; (-) Not real-world tested, needs good alignment
[171]	Good; 50% (unconditioned) and 37% (goal-conditioned) real robot success	High efficiency; uses internet videos only, no robot data; needs hand pose estimates	Moderate; Transformer-CVAE architecture with inverse kinematics controller. Requires camera calibration	Strong; hand trajectory mapping with camera transforms, IK	(+) No lab data, real-world tested; (-) Pose estimate noise, limited precision
[172]	High; SOTA on Meta-World/Robosuite (100% Lift, ~88% insert)	Very high; pretrains on 4500 video demos + 375 visuomotor, fine-tune with as few as 10 demos	High; Ddual Transformer (planner + executor) architecture with positional encoding and inverse dynamics training	Moderate; generalizes tasks, but assumes fixed robot morphology	(+) Multi-modal, few-shot; (-) High compute, needs curated data splits

Collectively, these works trace a clear trajectory: from hierarchical multimodal transformers, to prompt-based frameworks, to generative pretraining, and finally to large-scale VLA models that leverage internet video at unprecedented scales. The unifying theme is the shift from narrow, demonstration-driven learning toward architectures that internalize broad physical common sense, temporal dynamics, and general reasoning. These advances position multimodal foundation models not just as tools for policy learning, but as platforms for embodied intelligence with the capacity to adapt fluidly to new environments and tasks. Despite the many gains won by these models, Table 11 shows their

TABLE 11. Comparison of multi-modal transformers and large-scale foundation models for learning manipulation skills from human video.

Method	Task Performance	Sample Efficiency	Compute Cost	Embodi. Handling	Pros (+) / Cons (-)
[173]	SoTA on CALVIN: strong multi-stage, long-horizon manipulation	Moderate; leverages relabeled play/language data	Moderate-high (multimodal transformer + hierarchical structure)	Grounded language/vision	(+) Many tasks/flexible; (-) Model/training complexity
[175]	SoTA in visuo-motor multimodal tasks	Highly data-efficient via large-scale pretraining	High (transformer + large multimodal model)	Good; 3D scene reduces sim2real gap	(+) Strong zero-shot; (-) Expensive train/deploy
[176]	High; improves planning & execution	Efficient (retrieves knowledge instead of retraining)	Moderate (scene graph, LLM, simple policy)	Hierarchical narrows gap	(+) No retrain; (-) Needs scene graph/LLM reliability
[177]	SoTA on CALVIN and real robot tasks (94.9% success)	Very efficient (fine-tunes on 10% robot data)	High (GPT-style video transformer)	Good; finetunes after pretraining	(+) Best generalization/few-shot; (-) High model/training cost
[178]	SoTA on Meta-World, RLBench (both seen/unseen tasks)	High; learns from large unlabeled video, and requires few labeled demos	High (VQ-VAE, discrete diffusion, multi-stage training)	Visual token space bridges videos	(+) Leverages internet-scale data; (-) Compute heavy; limited real-robot
[101]	High; strong performance across simulation benchmarks	High; trained on diverse real-robot trajectories, human videos, and synthetic data	High; dual-system architecture with diffusion transformer	Excellent; tightly coupled VLA modules	(+) Open-source; (+) High data efficiency; (-) Complex model architecture
[100]	High; capable of completing 105 manipulation tasks with high success rate	High; pre-trained on 38 million text-video data	High; utilizes large multimodal model architecture	Good; supports various robot embodiments	(+) Strong generalization; (-) High compute requirements
[107]	High; outperforms state-of-the-art VLA models in real-world manipulation tasks	High; incorporates visual chain-of-thought reasoning for efficient task execution	Moderate; utilizes 7B parameter model	Good; trained on robot demonstrations and action-less videos	(+) Enhanced reasoning capabilities; (-) Requires fine-tuning for new tasks
[108]	High; excels in dexterous manipulation tasks	High; optimized for on-device processing with minimal computational resources	Moderate; designed for local inference on robots	Excellent; supports various robot embodiments	(+) On-device operation; (+) Adaptable to new tasks; (-) Limited to bi-arm robots

comparative extreme compute and data requirements make them unsuitable for low-resource settings.

IV. COMPARATIVE ANALYSIS BETWEEN THE APPROACHES

To provide a high-level synthesis of the literature, we present in Table 12 a macro-level comparison of the primary methodological subgroups discussed in this survey. While previous sections have detailed the specific technical approaches within each subgroup, this section aims to capture their overarching strengths, common pitfalls, bottlenecks, and representative applications.

The chosen metrics' main advantages, disadvantages, key bottlenecks, and reported applications reflect the most salient aspects influencing the applicability and impact of each subgroup in practical robot learning settings.

The comparative analysis reveals a field grappling with fundamental tensions between capability and practicality. At the foundational level, perception approaches illustrate this tension clearly: CNN-based methods offer computational efficiency but sacrifice fine-grained understanding, while pose/keypoint detection provides structured representations at the cost of robustness to estimation failures. This same pattern extends to domain bridging, where image translation prioritizes visual alignment over action precision, and context translation emphasizes environmental adaptability over training efficiency.

These trade-offs become more pronounced as approaches increase in sophistication. Reinforcement learning methods exemplify this progression, from computationally expensive visual RL to hierarchical frameworks that promise compositional learning but introduce complex decomposition challenges. Similarly, imitation learning has evolved from direct behavioral cloning toward meta-learning approaches that achieve impressive few-shot capabilities but require extensive pretraining regimes that paradoxically reduce their practical accessibility.

The field's response to these limitations through hybrid and multi-modal approaches reveals both promise and deeper challenges. Hybrid methods acknowledge that single

TABLE 12. Macro-level comparative analysis of video-based robot learning approaches.

Category	Advantages	Disadvantages	Key Bottlenecks	Representative Applications
Foundational Perception & Representation Methods	Interpretable representations; leverages large unlabeled datasets; real-time capable; enables structured motion capture; bridges visual domain gaps; adapts across environments	Limited to holistic features; dependent on estimation accuracy; translation artifacts; requires diverse contextual data; vulnerable to domain shift	Embodiment transfer, pose estimation robustness, visual realism vs. task fidelity, context generalization	Hand-object interaction, dexterous grasping, human-to-robot transfer, cross-domain manipulation, assembly tasks
Reinforcement Learning Approaches	Powerful for long-horizon tasks; handles complex manipulation; enables compositional learning; grounds policies in structured representations; supports multi-task generalization	High computational cost; complex architecture design; sensitive to reward engineering; challenging hierarchical decomposition; requires extensive interaction data	Feature representation quality, reward signal design, task decomposition, hierarchical learning stability	Object manipulation with dense rewards, multi-step assembly, hierarchical manipulation, tool use learning, compositional skill learning
Imitation Learning Approaches	Direct demonstration use; sample efficient; rapid skill acquisition; enables few-shot generalization; works with raw video data; leverages web-scale demonstrations	Distribution mismatch issues; requires extensive meta-training; embodiment mismatch challenges; sensitive to demonstration quality; limited generalization beyond training	Demonstration coverage, meta-learning stability, domain gap bridging, embodiment alignment, visual domain shift	One-shot imitation, few-shot manipulation, rapid task adaptation, human video imitation, cross-embodiment transfer
Hybrid Approaches	Combines RL/IL strengths; robust to demonstration noise; addresses cross-domain challenges; enables online refinement; flexible adaptation mechanisms	Increased design complexity; challenging balance optimization; complex pipeline design; limited real-world validation; task-specific solutions	Integration complexity, multi-objective learning, cross-domain alignment, multi-component integration, real-world deployment	Noisy demonstration learning, online policy refinement, cross-domain skill transfer, adaptive imitation, robust policy learning
Multi-Modal Approaches	Natural language task specification; unified representations; internet-scale pretraining; broad physical common sense; supports in-context learning; enables zero-shot generalization	Requires paired multimodal data; high model complexity; extremely high compute requirements; sensitive to language ambiguity; deployment challenges	Language-vision alignment, multimodal representation learning, computational scalability, data curation, model interpretability	Language-guided manipulation, trajectory prediction, general manipulation policies, zero-shot task execution, embodied reasoning

paradigms are insufficient, yet their attempt to combine multiple learning objectives often creates complex, task-specific solutions that resist broader generalization. Multi-modal approaches, particularly large-scale foundation models, represent the current apex of capability but exacerbate the accessibility problem by requiring computational resources beyond most practitioners' reach.

This evolution exposes a central paradox: the most capable methods are becoming increasingly inaccessible, while practical approaches face fundamental limitations in handling the embodiment transfer problem that persists across all categories. The field has not solved this core challenge but rather developed increasingly sophisticated ways to work around it, often at the cost of practical deployability.

The implications are clear: method selection cannot be driven solely by theoretical performance but must account for the specific constraints of computational resources, data availability, and deployment requirements. Rather than seeking universal solutions, the field is converging on a recognition that different approaches occupy distinct positions in a multi-dimensional trade-off space, making the matching of methods to specific scenarios the critical skill for practical robot learning applications.

V. OPEN-SOURCE TOOLS FOR VIDEO-BASED ROBOT MANIPULATION LEARNING

In this section, we discuss and provide an overview of open-source implementations, frameworks, tools, and datasets that constitute the foundation of modern video-based robot manipulation learning. The resources cataloged herein in table 13 to 18, span the entire learning pipeline, from foundational visual representation models that serve as the perceptual backbone of a system, to sophisticated end-to-

end VLA policies that map raw pixels and natural language commands directly to robot motor commands.

VI. CHALLENGES

Learning from video demonstrations presents several persistent challenges that cut across all major methodological subgroups. While remarkable progress has been made, limitations remain at both the data and model levels, and many of these challenges directly constrain the practical impact of state-of-the-art methods. This section identifies and discusses six key challenges in learning from video demonstrations, expanding on: (1) data availability and annotation, (2) domain shift and embodiment gap, (3) computational cost and scalability of learning architectures and resources, (4) model sample efficiency, (5) evaluation and benchmarking, and (6) causal reasoning and policy abstraction.

A. DATA AVAILABILITY AND ANNOTATION

The performance and generalization of video-based robot learning models remain highly dependent on the availability, diversity, and quality of demonstration data. As highlighted in our analysis, many approaches, especially feature extraction and imitation learning struggle when exposed to unfamiliar states or objects not seen during training. Large, balanced datasets are rare, and most existing datasets (e.g., EPIC-Kitchens [185], Something-Something [83]) are often domain-specific, unbalanced, or require expensive annotation. Specialized datasets such as Penn Action [186], HMDB51 [187], and MPII [188] are often limited in their diversity or designed for narrow, non-robotic tasks. Furthermore, several methods (e.g., Demo2Vec [52], ViP [141]) depend on expert-labeled demonstrations, further limiting their scalability in real-world scenarios. The community increasingly turns to scalable alternatives, such as leveraging uncurated internet videos, weak supervision, or self-supervised objectives, to address these data bottlenecks, but progress is ongoing.

B. DOMAIN SHIFT AND EMBODIMENT GAP

Domain shift remains a fundamental challenge across all subgroups. The disparity between human and robot domains, the so-called embodiment gap, often hinders the direct transfer of skills, as the visual appearance, dynamics, and even the action spaces differ substantially. Although some methods leverage domain-invariant or domain-adaptive features, as seen in hybrid, multi-modal, and image/context translation approaches, the problem is far from solved. Translation artifacts, imperfect pose estimation, and misaligned action representations can severely limit sim-to-real or human-to-robot transfer. Approaches like adversarial learning, keypoint-based transfer, and CycleGANs offer partial solutions, but robust, generalizable transfer remains elusive.

C. COMPUTATIONAL COST

Computational cost is an increasingly pressing concern, especially with the adoption of large-scale, multi-modal, and transformer-based architectures. While these models enable impressive performance, their high computational and memory requirements can limit real-world deployment, especially on resource-constrained robotic platforms. Many state-of-the-art methods (e.g., multi-modal and large-scale models III-E) demand substantial GPU resources for training and inference, hindering their use in real-time or edge settings. Furthermore, efficient scaling in both model size and data remains a bottleneck, with challenges in data collection, curation, and model parallelization.

D. SAMPLE EFFICIENCY

Although progress has been made, particularly with meta-learning, contrastive objectives, and one/few-shot imitation, the demand for large volumes of video data remains a core challenge, especially for RL-based and high-capacity models. Many algorithms still require thousands of demonstrations or millions of interaction steps, which can be prohibitive in real-world settings. While approaches like feature extraction III-A1, goal-conditioned RL III-B, and some hybrid models III-D are relatively more data-efficient, the quest for robust learning from minimal or weakly-labeled data is ongoing. Bridging this gap is crucial for broadening the applicability of video-based robot learning in low-data regimes.

E. EVALUATION METRICS AND BENCHMARKING

A critical challenge identified across the literature is the lack of standardized evaluation metrics and benchmarking protocols for video-based robot learning. Unlike computer vision or natural language processing, where large-scale public benchmarks drive progress, robotics evaluations are often fragmented and task-specific, relying on human judgment, custom setups, or bespoke datasets. This fragmentation complicates fair comparisons between approaches and slows the pace of reproducible progress. There is a growing need for community-driven, standardized benchmarks and clearly defined metrics that capture not only task success but also generalization, robustness, and sim-to-real performance.

F. CAUSAL REASONING AND POLICY ABSTRACTION

A final major bottleneck, as surfaced in our comparative analysis, is the limited capacity of current methods to perform causal reasoning and high-level policy abstraction from video data. Most approaches focus on pattern recognition, goal inference, or direct imitation (see Section III), rarely incorporating causal structure or relational reasoning about actions and outcomes. As a result, models may lack robustness when faced with novel tasks or environments that require understanding of underlying cause-and-effect relationships.

TABLE 13. Foundational visual representation learning resources.

Resource Name	Primary Function	Key Features	Open-Source Link
Time-Contrastive Networks (TCN) [35]	Self-Supervised Representation Learning	• Learns viewpoint-invariant features from multi-view video • Uses a time-contrastive triplet loss • Focuses on temporal dynamics over static appearance	https://github.com/kekeblom/tcn
Spatiotemporal Contrastive Video Representation Learning (CVRL) [44]	Self-Supervised Representation Learning	• Learns spatiotemporal features via contrastive loss • Employs temporally consistent spatial augmentations • Outperforms ImageNet pre-training on video tasks	https://github.com/tensorflow/models/tree/master/official/
Dense Predictive Coding (DPC) [39]	Self-Supervised Representation Learning	• Predicts future latent representations, not pixels • Learns dense spatio-temporal block embeddings • Uses curriculum learning to predict further in time	https://github.com/TengdaHan/DPC
Masked Visual Pre-training (MVP) [45]	Self-Supervised Representation Learning	• Extends Masked Autoencoders (MAE) to robotics • Pre-trains on large image/video datasets • Frozen encoder serves as perception module for control	https://github.com/ir413/mvp
R3M [47]	Universal Visual Representation	• Pre-trained on Ego4D human video dataset • Combines time-contrastive and video-language learning • Serves as a frozen perception module for manipulation	https://sites.google.com/view/robot-r3m

TABLE 14. 3D hand & body modeling resources.

Resource Name	Primary Function	Key Features	Open-Source Link
SMPL [76]	Parametric 3D Body Model	• Learned from thousands of 3D body scans • Low-dimensional shape and pose parameterization • Compatible with standard graphics engines	https://smpl.is.tue.mpg.de/
MANO [75]	Parametric 3D Hand Model	• Specialized parametric model for the human hand • Integrates with SMPL to form SMPL+H • Enables detailed, articulated hand modeling	https://mano.is.tue.mpg.de/
FrankMocap [73]	Monocular Motion Capture System	• Real-time 3D hand and body motion capture from single RGB video • Leverages SMPL-X for unified parametric output • Enables data extraction from in-the-wild videos	https://github.com/facebookresearch/frankmocap

Bridging this gap will likely require new architectures and training paradigms that integrate causal inference, model-based reasoning, or neuro-symbolic methods, as well as datasets that explicitly capture causal interactions.

TABLE 15. Affordance & interaction resources.

Resource Name	Primary Function	Key Features	Open-Source Link
AffordanceNet [54]	Affordance Detection	• End-to-end object and affordance detection from RGB-D • Uses a multi-task, two-branch architecture • Segments pixels into functional categories (e.g., "grasp")	https://github.com/wliu88/affordance_net
Demo2Vec [52]	Affordance Reasoning	• Learns video embeddings to reason about affordances • Predicts interaction heatmaps and action labels on a target image • Trained on YouTube product review videos	https://sites.google.com/view/demo2vec/
Vision-Robotics Bridge (VRB) [53]	Affordance Grounding	• Learns agent-agnostic affordances from human videos • Predicts contact heatmaps and post-contact trajectories • Integrates with multiple robot learning paradigms	https://robo-affordances.github.io/

TABLE 16. Vision-language-action (VLA) policies.

Resource Name	Primary Function	Key Features	Open-Source Link
RT-1 [14]	Transformer-Based Policy	• End-to-end Transformer for multi-task control • Tokenizes images, language, and actions • Trained on 130k+ real-world robot trajectories	https://github.com/google-research/robotics_transformer
RT-2 [15]	Vision-Language-Action Model	• Fine-tunes web-scale VLMs for robotic control • Represents robot actions as text tokens • Transfers semantic knowledge from web data to robotics	https://github.com/kyegomez/RT-2
CLIPort [12]	Language-Conditioned Imitation	• Two-stream architecture: semantic (CLIP) and spatial (Transporter) • Combines "what" (language) and "where" (affordance) • Generalizes to unseen objects without explicit detectors	https://cliport.github.io/
VIMA [175]	Multimodal Prompting Agent	• Generalist Transformer agent for diverse tasks • Unifies task specification via multimodal prompts (text + vision) • Trained on 600k+ expert trajectories in VIMA-Bench	https://vimalabs.github.io/

These challenges, distilled from a broad cross-section of the literature and subgroup analysis, highlight that progress in video-based robot learning is not uniform; each methodological paradigm offers distinct strengths but also faces recurring limitations. Addressing these bottlenecks, particularly in data transfer, scalability, and abstraction, will be key to achieving robust, generalizable, and efficient robot learning from videos in the future.

VII. FUTURE OUTLOOK

As discussed in previous sections, various approaches have been proposed for learning manipulation skills through video demonstrations. We also explored the challenges and limitations of learning from videos. This section will spotlight several promising but relatively underexplored areas in video-based learning research. These areas include data efficiency, interactive and active learning, multi-task

TABLE 17. Datasets & simulators.

Resource Name	Primary Function	Key Features	Open-Source Link
Open X-Embodiment (OXE) [16]	Cross-Embodiment Dataset	• Unifies 60+ datasets from 22 robot embodiments • 1M+ real robot trajectories in standardized RLDS format • Enables training of generalist "X-robot" policies	https://github.com/google-deepmind/open_x_embodiment
DROID [88]	In-the-Wild Manipulation Dataset	• 76k+ trajectories from 50+ global users • Collected in 564 diverse, unstructured scenes • Designed for studying real-world generalization	https://droid-dataset.github.io/
BridgeData V2 [17]	Multi-Task Manipulation Dataset	• 60k+ trajectories on a low-cost WidowX robot • Spans 24 environments and 13 skills • Includes language and goal-image conditioning	https://rail-berkeley.github.io/bridgedata/
RoboNet [18]	Multi-Robot Video Dataset	• 15M video frames from 7 different robot platforms • Early large-scale effort to share robotic experience • Designed for learning generalizable vision-based models	https://www.robonet.wiki/
RH20T [19]	Contact-Rich Manipulation Dataset	• 110k+ sequences focusing on contact-rich skills • Rich multi-modal data (vision, force, audio, action) • Includes paired human demonstration for each robot sequence	https://rh20t.github.io/
Isaac Sim [179]	GPU-Based Physics Application	• High-performance simulator for massively parallel RL • Enables training policies directly on GPU • Developed by NVIDIA	https://developer.nvidia.com/isaac/sim
MuJoCo [180]	Physics Engine	• Fast and accurate physics simulation • Widely used for robotics, biomechanics, and RL research • Features optimization-based contact dynamics	https://mujoco.org/
PyBullet [181]	Physics Simulator	• Python module for the Bullet physics engine • Provides robotics simulation and RL environments • Open-source and widely accessible	https://pybullet.org/

learning architectures, integration of causal reasoning, and the development of evaluation metrics and benchmarks.

A. TACKLING DATA EFFICIENCY AND AVAILABILITY

Addressing the challenges discussed in Section VI-A regarding data availability and annotation is crucial. Works such as [14], [15], [16], [17], [18], [19] have made dedicated efforts to collect extensive data for training robots in various skills. While these endeavors contribute valuable datasets, video demonstrations offer unique advantages compared to task-specific datasets. Despite the risk of introducing biases, video data provides a more unbiased and diverse representation of real-world scenarios, fostering improved generalization. Additionally, videos capture realistic dynamics and environmental variability, enabling models to better handle uncertainties and variations encountered in real-world scenarios.

As discussed in Section VI-A, poor generalization may lead robots to struggle with tasks in states not adequately

TABLE 18. Core libraries.

Resource Name	Primary Function	Key Features	Open-Source Link
OpenCV [182]	Computer Vision Library	• De facto standard library for real-time image/video processing • Provides a vast suite of foundational CV algorithms • Underpins most vision-based robotics research	https://opencv.org/
MediaPipe [183]	Perception Pipeline Framework	• Cross-platform framework for applied ML pipelines • Offers pre-trained models for pose, hand, and face tracking • Used as off-the-shelf perception components	https://developers.google.com/mediapipe
OpenPose [184]	2D Pose Estimation	• Real-time multi-person 2D keypoint detection • Detects body, hand, foot, and facial keypoints • Widely used for human activity analysis	https://github.com/CMU-Perceptual-Computing-Lab/openpose

covered during training. Generalization is a pervasive topic in deep learning, and several works, including [14], [141], [150], propose techniques for learning from videos to enhance model generalization across diverse robots, tasks, and states. However, current approaches still have limitations in their generalization, particularly to tasks not recorded in the video demonstration data. Future work should focus on addressing this limitation by identifying intuitive methods to ensure not only generalization but also quick adaptation of these models to changing tasks and environments.

B. IMPROVING DATA ANNOTATION THROUGH ACTIVE LEARNING

Ensuring high-quality data for training models is crucial. Active learning strategies, such as those proposed by [189] and [190], empower robots to strategically select informative data points, optimizing the learning process by Intelligently querying labels on challenging instances. This approach reduces the reliance on extensive labeled datasets while maintaining or improving performance.

Current approaches often passively observe large demonstration datasets [83], [185], which can be expensive to scale up. Integrating active physical trials on real robots alongside video data observation combines the strengths of imitation learning and embodied reinforcement learning. This approach helps bridge the reality gap by incorporating interactions in physical environments, dynamic feedback, and recalibration of visual interpretations. It allows robots to adapt to environmental changes, providing signals when contexts shift and offering opportunities for adaptation. Passive video datasets often lack diversity across potential deployment environments, making the inclusion of physical interactions valuable. This approach not only enables the robot to learn what works but also what doesn't work well in different situations.

While studies like [156] and [149] have proposed related techniques, they have not thoroughly explored grounding policies learned from video data in physical environments. Future works could benefit from addressing these points.

C. TACKLING DOMAIN SHIFT

Future work should focus on addressing the persistent challenge of domain shift between human and robot domains. Works like [191] addressed domain shift in the context of appearance changes in outdoor robotics with adversarial domain adaptation, while [192] presented a survey in learning for robot decision making under distribution shift. Advanced domain adaptation techniques, including adversarial training and meta-learning, could create more robust and generalizable models. Multi-modal learning strategies that incorporate additional sensory inputs may reduce reliance on visual domain translation. Sim-to-real transfer methods and continual learning paradigms offer promising avenues for improving domain adaptation. Investigating attention mechanisms, unsupervised techniques, and transformer-based architectures could yield more effective domain-invariant features. Additionally, exploring causal reasoning and few-shot learning approaches may enhance the efficiency of skill transfer from human demonstrations to robotic applications. By pursuing these strategies, future research can work towards mitigating the impact of domain shift and improving the effectiveness of learning from videos for robotic manipulation tasks.

D. INTRODUCING EMERGING TECHNIQUES AND ARCHITECTURES

The techniques employed in the studies outlined in Section III predominantly rely on single modalities or involve single-task architectures. Recent research emphasizes the efficacy of learning from multiple tasks and modalities [193], [194], [195], [196]. Studies like [141] and [150] discussed

in Section III underscore the effectiveness of multi-task learning for acquiring robot manipulation skills from videos. Challenges inherent in multi-tasking, extensively explored in studies like [197], become more pronounced due to the varying optimization constraints between predicting actions from fixed videos and closed-loop control problems.

Furthermore, the pursuit of learning manipulation skills from video has spurred the development of increasingly sophisticated generative architectures. These models have moved beyond simple regression or classification to generate complex, high-dimensional outputs like action trajectories and future video frames. The leading emerging architectural paradigms include diffusion models and world models.

Diffusion generative models have rapidly emerged as a dominant force in robotics, prized for their ability to model complex data distributions and their robustness in high-dimensional spaces [198]. Their application to visuomotor control represents a significant step forward from prior generative approaches. Diffusion Probabilistic Models (DMs) operate on a simple yet powerful principle, executed in two stages [198]. The first is a fixed forward process, where Gaussian noise is progressively added to a data sample (e.g., an image or an action trajectory) over a series of timesteps, gradually corrupting it into pure noise. The second stage is a learned reverse process, where a neural network, typically a U-Net or a Transformer, is trained to reverse this noising process. By learning to predict and remove the noise at each step, the model can start with a random noise tensor and iteratively denoise it to generate a new, high-fidelity sample from the original data distribution [199].

A key advantage of this framework for robotics is its inherent capacity to model multi-modal distributions [198]. Many manipulation tasks do not have a single correct solution; there can be multiple valid trajectories to pick up an object or open a drawer. Traditional methods that predict a single, unimodal output (e.g., by minimizing mean squared error) tend to average these possibilities, resulting in mediocre or invalid actions. Diffusion models, by contrast, can capture the full distribution of successful behaviors, allowing them to generate diverse and plausible action sequences at inference time [199].

This capability has been elegantly harnessed in the ''Diffusion Policies'' framework for imitation learning [199]. In this paradigm, the model learns a policy that generates robot actions directly from visual observations. The policy is a diffusion model trained to denoise an action trajectory conditioned on the current visual state of the environment. At each step of the reverse process, the model refines its prediction of the entire action sequence, leading to temporally coherent and precise behaviors [200].

While diffusion policies excel at learning direct mappings from observations to actions, a parallel and complementary approach involves learning a predictive model of the world itself. These ''world models'' enable an agent to plan and reason by simulating the future consequences of its actions internally, a process often referred to as ''imagining'' [112]. A world model learns the transition dynamics of an environment, formally represented as the probability distribution $p(s_{t+1}|s_t, a_t)$, where s is the state and a is the action. By repeatedly applying this learned model, an agent can forecast entire trajectories of future states that would result from a sequence of actions. This predictive capability is the foundation for model-based planning, where the agent can search for the optimal action sequence within its learned model before executing it in the real world. This approach can be significantly more sample-efficient than model-free methods, which must learn through extensive trial-and-error in the physical environment [201].

A powerful and intuitive instantiation of a world model is a video prediction model. Here, the state s is represented by an image or a sequence of images, and the world model learns to generate future video frames conditioned on a sequence of actions. Mani-WM [202] is a prime example of this approach, leveraging a diffusion transformer to generate high-resolution, long-horizon videos of a robot arm executing a specified action trajectory. The model employs a novel frame-level conditioning technique to ensure precise temporal alignment between the generated frames and the input actions. The resulting learned model serves as a high-fidelity, interactive simulator. This allows for downstream applications like policy evaluation and model-based planning to be conducted entirely with the generative model, mitigating the cost, safety concerns, and labor associated with extensive real-world robot rollouts [202].

E. ROBUST EVALUATION METRICS AND BENCHMARKS

The rapid pace of architectural innovation necessitates equally rigorous and standardized evaluation methodologies. A model's claimed contributions are only as strong as the benchmarks used to validate them. We discussed below the most prominent benchmarks in video-based manipulation, analyzing their strengths, intended research focus, and, critically, their limitations, particularly concerning the evaluation of causal understanding.

1) RLBENCH: A TESTBED FOR BROAD SKILL ACQUISITION

RLBench stands as a cornerstone for evaluating the breadth and generalization of manipulation skills [203]. Its primary contribution is a massive and diverse suite of 100 unique, hand-designed tasks simulated in the V-REP (now CoppeliaSim) environment. These tasks span a wide spectrum of difficulty, from simple behaviors like reaching a target or opening a door to complex, multi-stage sequences like opening an oven and placing a tray inside.

A defining feature of RLBench is its rich multi-modal observation space, providing agents with RGB-D images from both a static over-the-shoulder camera and an eye-in-hand camera, as well as proprioceptive data like joint angles and torques. Perhaps its most unique and powerful feature is the provision of a virtually infinite supply of expert

demonstrations for every task. These demonstrations are generated via motion planners operating on pre-defined waypoints, enabling a wide range of research in imitation learning and reinforcement learning that leverages expert data [203]. The benchmark is explicitly designed to push research in multi-task learning, meta-learning, and, in particular, few-shot learning.

2) CALVIN: THE STANDARD FOR LONG-HORIZON LANGUAGE GROUNDING

While RLBench tests the breadth of skill acquisition, CALVIN (Composing Actions from Language and Vision) is the de facto standard for evaluating the depth of long-horizon, compositional reasoning [174]. CALVIN is an open-source simulated benchmark, built in PyBullet, designed to develop and test agents that can solve complex manipulation tasks specified solely by natural language instructions. A single agent must learn to understand and execute a sequence of commands, such as ''open the drawer, pick up the blue block, push the block into the drawer, open the sliding door'' [174].

CALVIN's key contribution is its focus on long-horizon problems and language-based generalization. The benchmark includes four distinct environments (A, B, C, D) with shared structure but different visual textures and object layouts, allowing for rigorous testing of zero-shot generalization to novel scenes. The provided dataset is not a set of isolated, task-specific demonstrations, but rather hours of unstructured ''play data,'' from which task sequences are procedurally labeled. This setup mimics a more realistic learning scenario where an agent must discover skills from continuous interaction data. Evaluation is performed on the agent's ability to generalize to novel language instructions and to complete long sequences of tasks, which is highly challenging as it requires the agent to robustly transition between different subgoals without compounding errors. Its status as a challenging and well-defined testbed has made it the proving ground for state-of-the-art models like VidMan [204].

3) ROBOTUBE: BRIDGING THE HUMAN-TO-ROBOT GAP

A central goal of the field is to enable robots to learn directly from observing humans. RoboTube is a benchmark designed specifically to facilitate research toward this goal [205]. It directly addresses the limitations of prior video datasets, which often lack task complexity or relevance to household robotics. The RoboTube dataset consists of 5,000 high-quality, multi-view RGB-D video demonstrations of humans performing a variety of complex household tasks. These include the manipulation of not just rigid objects, but also articulated objects (drawers, cabinets), deformable objects (cloth), and granular materials (pouring).

The most significant feature of RoboTube is its ''simulated twin'' environment, RT-sim [205]. The objects and scenes from the real-world videos have been meticulously 3D-scanned to create photo-realistic, physically accurate digital counterparts in simulation. This unique pairing of a real human video dataset with a high-fidelity simulated testbed is invaluable. It provides a controlled, reproducible platform for researchers to develop and benchmark algorithms for key challenges like sim-to-real transfer, representation learning from human video, and self-supervised reward learning, with the confidence that models validated in RT-sim have a clear path to deployment on a real robot. RoboTube aims to democratize research in this area by lowering the barrier to entry and providing a standardized platform for comparing different approaches to learning from human videos.

The design of a benchmark implicitly steers the research priorities of the community. The existence of CALVIN has catalyzed a wave of innovation in long-horizon, language-conditioned policies [174], while RLBench has standardized the evaluation of few-shot and multi-task learning [203]. A critical analysis of these leading benchmarks, however, reveals a significant gap: none of them are explicitly designed to evaluate an agent's causal understanding of the world. An agent can achieve a high success rate on a CALVIN task by mastering the statistical correlations present in the massive demonstration dataset. It might learn that pushing a red button is followed by a light turning on, but it may not have learned the underlying causal link. This purely correlational policy would fail if the button's function were rewired, an object's physical properties were altered, or an unobserved confounder were introduced.

The MVP (Minimal Video Pairs) [206] benchmark, developed for visual question answering, offers a blueprint for how to address this gap. MVP consists of pairs of videos that are minimally different, often with a single changed detail, accompanied by the same question but with opposite answers. This design forces a model to move beyond superficial cues and engage in deeper reasoning to arrive at the correct answer. This principle must be extended to robotic manipulation to properly evaluate the benefits of causal models. A novel and impactful contribution would be the development of a new evaluation protocol, which could be termed ''Causal-CALVIN'' or ''Interventional RLBench.'' In this protocol, an agent would first be trained on the standard benchmark dataset. Then, its generalization and robustness would be tested on a suite of evaluation tasks where the underlying causal structure of the environment has been perturbed. For example, the mass or friction of an object could be significantly changed, the causal link between a switch and a light could be broken or rewired to a different switch, or a previously free-sliding drawer could be made to stick. A purely correlational model, having overfitted to the statistics of the training environment, would be expected to fail catastrophically. In contrast, a model equipped with an accurate causal model of the world should be able to either adapt its policy to the new dynamics or at least recognize that its model of the world is no longer valid, enabling more robust and intelligent failure recovery. This provides a concrete, quantitative methodology for measuring the tangible benefits of causal reasoning, moving evaluation beyond simple task

success rates to a more meaningful assessment of physical understanding.

Table 19 provides a strategic overview of these benchmarks, enabling the selection of appropriate evaluation platforms to highlight different facets of a proposed model's performance.

F. INTEGRATION OF CAUSAL REASONING

To build robots that can operate robustly and adaptively in the open world, it is essential to move beyond learning statistical correlations and toward models that capture the underlying causal structure of their environment. Consequently, below we provide a detailed investigation into the principles and methods of causal reasoning as they apply to video-based manipulation, establishing the foundation for a research agenda centered on this critical capability.

The vast majority of modern machine learning models, including deep neural networks, are powerful function approximators trained to minimize a loss function on an independently and identically distributed (i.i.d.) dataset. This process enables them to excel at learning complex correlations within the training data. However, correlation does not imply causation. This fundamental limitation is the root cause of their brittleness when deployed in the real world, which is inherently non-stationary and subject to constant distribution shifts [207]. A robot trained in a lab may learn a spurious correlation between the color of a block and its weight, and will fail when presented with a block of a different color.

Causal models offer a principled escape from this trap [208]. By aiming to represent the actual data-generating processes, causal models provide a foundation for true generalization. A causal model understands that an object's mass, not its color, determines the force required to lift it. This understanding allows for robust performance even when encountering novel objects and conditions. For robotics, the promise of causality is threefold: robustness to environmental changes, generalization to novel scenarios, and the ability to perform counterfactual reasoning (e.g., ''what would have happened if I had pushed the object instead of grasping it?''). Which is the bedrock of intelligent planning and decision-making [208].

1) METHODS FOR CAUSAL DISCOVERY FROM VISUAL DATA

Causal discovery is the process of inferring the causal structure (typically represented as a Directed Acyclic Graph, or DAG) from data [209]. In robotics, this means discovering the cause-and-effect relationships between objects, actions, and environmental variables from sensory inputs like video. These methods can be broadly categorized into observational and interventional approaches.

- **Observational Approaches:** Observational methods attempt to uncover causal structure from passively collected data, without actively manipulating the system. These approaches typically fall into two families: constraint-based and score-based methods. Constraint-based algorithms, like the PC algorithm [209], work by performing a series of conditional independence tests on the data to prune edges from a fully connected graph. Score-based methods define a scoring function (e.g., Bayesian Information Criterion) that measures how well a given graph structure fits the data and then search the space of possible graphs for the one with the best score. Applying these methods to high-dimensional video data requires specialized architectures. The Visual Causal Discovery Network (V-CDN) [208] is a seminal work in this area. V-CDN is an end-to-end model that learns to discover causal relationships in physical systems directly from video. It consists of three key modules:
  1) Perception Module: Extracts an unsupervised, temporally consistent keypoint representation of objects in the scene from raw images. These key points serve as the variables in the causal graph.
  2) Inference Module: Observes the dynamics of these keypoints over a short video sequence and infers a latent causal graph, determining which keypoints are causally related (e.g., connected by a spring or a rigid rod).
  3) Dynamics Module: A graph neural network that takes the inferred causal graph as input and learns to predict the future evolution of the system.

  A crucial aspect of V-CDN is its assumption that the training data, while passively observed, is sourced from a variety of configurations and environmental conditions. This is treated as data from ''unknown interventions'' on the system [208]. For example, by observing videos of systems with different numbers of objects or different spring constants, the model can disambiguate direct causal links from mere correlations and identify the correct underlying causal graph without requiring explicit labels for the interventions performed [208].
- **Interventional Approaches:** While observational methods are powerful, the gold standard for establishing causality is intervention: the act of actively manipulating a variable, denoted as $do(X = x)$, and observing the effect on the system [210]. Interventions break potential confounding pathways and provide unambiguous evidence of cause-and-effect relationships. Robots, as physically embodied agents, are uniquely positioned to perform such interventions, making them ideal platforms for active causal discovery [211].

  A powerful demonstration of this principle is SCALE (Skills from CAusal LEarning) [212]. SCALE addresses the problem of learning manipulation skills that generalize across different contexts. Instead of learning a single, complex policy over a high-dimensional state space, SCALE uses a simulator as a ''causal reasoning engine'' to perform targeted interventions. For a given task, it systematically perturbs context variables

TABLE 19. A comparison of benchmarks for robotics.

Benchmark	Primary Focus	# of Tasks	Data Provided	Key Evaluation Metric(s)	Simulation Environment	Suitability for Causal Evaluation
CALVIN	Long-horizon, language-conditioned, compositional tasks.	34	~24 hours of unstructured, teleoperated "play" data; language annotations for ~1% of data.	Long-Horizon Multi-Task Language Control (LH-MTLC) success rate.	PyBullet	Low (as is); High (with modification). The current setup rewards correlational learning. A "Causal-CALVIN" variant with perturbed physics/mechanisms would be required to test causal understanding.
RLBench	Few-shot, meta-, and multi-task learning; broad skill acquisition.	100	Infinite supply of motion-planned expert demonstrations for every task; multi-modal observations (RGB-D, proprioception).	Few-shot task success rate.	CoppeliaSim (V-REP)	Low (as is); High (with modification). The diversity of tasks is high, but the physics within each task is fixed. An "Interventional RLBench" with variations in physical properties (mass, friction, etc.) would be needed.
RoboTube	Learning from human video demonstrations; human-to-robot transfer.	~50	5,000 multi-view RGB-D videos of human demos; a "simulated twin" (RT-sim) with photo-realistic assets.	Policy success rate in RT-sim after training on human videos; sim-to-real transfer success.	Custom (RT-sim)	Moderate. The paired real/sim setup is ideal for studying the transfer of causal models. The diversity of object types (deformable, granular) provides a rich testbed for models of complex causal interactions.

(e.g., object positions, sizes, masses) and observes whether the intervention affects the task outcome. This process allows it to identify the minimal subset of context variables that are causally relevant for success. It then learns a ''compressed'' skill or option that is conditioned only on this small set of causal variables, ignoring all spurious features. This results in policies that are dramatically more sample-efficient and exhibit superior sim-to-real transfer, as they are not distracted by irrelevant, correlational features of the training environment [212].

Building on this, the Causal Robot Discovery (CRD) framework proposes a continual, online approach to causal learning [213]. The robot begins by building an initial causal model from passive observation. It then analyzes this model to identify the most uncertain or unreliable links (e.g., those with high p-values from conditional independence tests). Based on this uncertainty, the robot plans and executes its next set of interventions specifically to gather data that will maximally resolve this uncertainty [213]. This creates an efficient, self-improving feedback loop where the robot actively seeks the most informative data to refine its causal understanding of the world, making it particularly well-suited for resource-constrained robotics applications [213].

2) CAUSAL REPRESENTATION LEARNING (CRL)

The effectiveness of any causal discovery method depends on the variables over which it operates. Causal Representation Learning (CRL) is an emerging field that aims to bridge the gap between low-level sensory data (like pixels) and high-level causal variables [214]. The goal of CRL is to learn a mapping from high-dimensional observations to a low-dimensional latent space where the axes of the representation correspond to the independent causal mechanisms of the world [207]. For example, an ideal causal representation of a scene would disentangle factors like object identity, pose, lighting, and background into separate, independently controllable latent variables. Such a representation is inherently more compositional and generalizable, as the model can reason about and manipulate these factors independently, a key requirement for advanced robotics and embodied AI [214].

The motivation for incorporating causality into robotics must be driven by tangible, pragmatic benefits that improve manipulation performance. It is not merely a pursuit of philosophical purity or interpretability. The evidence from recent work clearly demonstrates that causal reasoning is a practical tool for building more efficient, robust, and intelligent robots.

The most direct evidence comes from the SCALE framework, which links causal discovery directly to improved policy learning [212]. By using interventions in a simulator to identify the true causal drivers of task success, SCALE learns a compressed policy that is conditioned only on relevant variables. This policy is not only more sample-efficient to train but also more robust to spurious correlations in the environment. When transferred to the real world, this causally-informed policy succeeds where a standard,

TABLE 20. Comparison of causal discovery methods in robotics.

Method	Core Principle	Data Requirement	Key Application/Benefit	Citation
V-CDN	End-to-end discovery of latent causal graphs from video via keypoint extraction and graph inference.	Observational (passive videos from diverse, "unknown" interventional settings).	Enables learning of predictive dynamics models that can perform counterfactual reasoning and extrapolate to unseen system configurations.	9
SCALE	Active interventional discovery of causally relevant features for a task to learn a compressed, robust skill.	Interventional (requires a simulator or "causal reasoning engine" to perform targeted interventions).	Improves sample efficiency and sim-to-real transfer by learning policies that are robust to spurious correlations.	11
CRD	Continual, online refinement of a causal model by using model uncertainty to guide the next set of active interventions.	Hybrid (Observational + Interventional). The robot actively collects interventional data to improve its model.	Enables efficient causal discovery on resource-constrained robots by creating a self-improving, active learning loop.	43
F-PCMCI	An efficient, filtered version of the PCMCI algorithm for causal discovery from time-series data.	Observational (time-series data).	Designed for fast and accurate causal analysis in real-time robotics applications, such as modeling human-robot interaction.	45

correlational policy fails, providing a clear demonstration of how causal feature selection enhances generalization and sim-to-real transfer.

Furthermore, causal knowledge can make the entire learning process more efficient. A robot that understands the causal structure of its environment can guide its exploration more intelligently. Instead of exploring randomly, an RL agent can use its causal model to prioritize actions that are most likely to influence task-relevant variables, dramatically reducing the number of samples needed to learn an effective policy [210]. The CRD framework operationalizes this by using the uncertainty in its current causal model to actively plan the most informative interventions, creating a highly efficient data collection loop [213].

Finally, a validated causal model unlocks the ability to perform counterfactual reasoning, which is the foundation of robust, deliberative planning. The model can simulate the outcomes of actions it has never taken in situations it has never seen, allowing it to plan for novel circumstances and recover from failures [208]. For instance, it can be used to analyze why a task failed by tracing back the chain of causal events that led to the undesirable outcome, enabling more sophisticated error diagnosis and correction [212].

Table 20 provides a taxonomy of key causal discovery methods relevant to robotics, organizing the literature into a coherent framework and highlighting their practical applications.

VIII. CONCLUSION

In this survey, we reviewed the emerging paradigm of robot learning for manipulation skills by leveraging abundantly available uncurated videos. Learning from video data allows for better generalization, reduction in dataset bias, and cutting down the costs associated with obtaining well-curated datasets. We began by outlining and discussing the essential components required for learning from video data and some current large-scale datasets and network architectures proposed for robot learning.

We surveyed techniques spanning representational, reinforcement, imitation, hybrid, and multimodal learning approaches for learning from demonstration videos in an end-to-end or modular manner. Analysis was provided around representations, sample efficiency, interpretability, and robustness for categories like pose estimation, image translation, and vision-language approaches. The benefits highlighted include generalization beyond controlled environments, scalability through abundant supervision, and avoiding biases coupled with human dataset curation. We also discussed evaluation protocols, sim-to-real challenges, and interactive learning as augmentations to pure video-based learning.

In conclusion, while still a nascent research direction, robot learning from online human videos shows immense promise in overcoming key data challenges prevalent in other supervised manipulation learning paradigms. If open challenges around dynamics, long-horizon understanding, absence of consistent and objective evaluation and benchmarking protocols, and sim-to-real transfer are systematically addressed, video-based learning can provide a scalable, economical, and generalizable pathway for robot acquisition of intricate real-world manipulation skills.

REFERENCES

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[2] S. James, M. Bloesch, and A. J. Davison, "Task-embedded control networks for few-shot imitation learning," in *Proc. Conf. robot Learn.*, 2018, pp. 783–795.

[3] S. Kadam and V. G. Vaidya, "Review and analysis of zero, one and few shot learning approaches," in *Proc. 18th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, 2019, pp. 100–112.

[4] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," in *Proc. Conf. robot Learn.*, 2017, pp. 357–368.

[5] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, and S. Levine, "One-shot imitation from observing humans via domain-adaptive meta-learning," 2018, *arXiv:1802.01557*.

[6] A. H. Qureshi, Y. Miao, A. Simeonov, and M. C. Yip, "Motion planning networks: Bridging the gap between learning-based and classical motion planners," *IEEE Trans. Robot.*, vol. 37, no. 1, pp. 48–66, Feb. 2021.

[7] A. Fishman, A. Murali, C. Eppner, B. Peele, B. Boots, and D. Fox, "Motion policy networks," in *Proc. Conf. Robot Learn.*, 2022, pp. 967–977.

[8] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," 2017, *arXiv:1709.10087*.

[9] H. Zhu, A. Gupta, A. Rajeswaran, S. Levine, and V. Kumar, "Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3651–3657.

[10] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-shot task generalization with robotic imitation learning," in *Proc. Conf. Robot Learn.*, 2022, pp. 991–1002.

[11] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "DexMV: Imitation learning for dexterous manipulation from human videos," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 570–587.

[12] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learn.*, 2021, pp. 894–906.

[13] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," 2018, *arXiv:1805.01954*.

[14] A. Brohan et al., "RT-1: Robotics transformer for real-world control at scale," 2022, *arXiv:2212.06817*.

[15] A. Brohan et al., "RT-2: Vision-language-action models transfer Web knowledge to robotic control," in *Proc. 7th Annu. Conf. Robot Learn.*, 2023, pp. 1–7.

[16] A. O'Neill et al., "Open X-embodiment: Robotic learning datasets and RT-X models," 2023, *arXiv:2310.08864*.

[17] H. Walke, K. Black, A. Lee, M. Jin Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "BridgeData v2: A dataset for robot learning at scale," 2023, *arXiv:2308.12952*.

[18] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "RoboNet: Large-scale multi-robot learning," 2019, *arXiv:1910.11215*.

[19] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 653–660.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[21] A. Gupta, A. Murali, D. Gandhi, and L. Pinto, "Robot learning in homes: Improving generalization and reducing dataset bias," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9094–9104.

[22] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, and P. Bojanowski, "Self-supervised pretraining of visual features in the wild," 2021, *arXiv:2103.01988*.

[23] Y. Tian, O. J. Hénaff, and A. V. D. Oord, "Divide and contrast: Self-supervised learning from uncurated data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10043–10054.

[24] M. Caron, P. Bojanowski, J. Mairal, and A. Joulin, "Unsupervised pre-training of image features on non-curated data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2959–2968.

[25] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9879–9889.

[26] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto, "From play to policy: Conditional behavior generation from uncurated robot data," 2022, *arXiv:2210.10047*.

[27] M. Q. Mohammed, L. C. Kwek, S. C. Chua, A. Al-Dhaqm, S. Nahavandi, T. A. E. Eisa, M. F. Miskon, M. N. Al-Mhiqani, A. Ali, M. Abaker, and E. A. Alandoli, "Review of learning-based robotic manipulation in cluttered environments," *Sensors*, vol. 22, no. 20, p. 7938, Oct. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/20/7938

[28] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auto. Syst.*, vol. 57, no. 5, pp. 469–483, May 2009.

[29] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annu. Rev. Control, Robot., Auto. Syst.*, vol. 3, no. 1, pp. 297–330, May 2020.

[30] A. Lobbezoo, Y. Qian, and H.-J. Kwon, "Reinforcement learning for pick and place operations in robotics: A survey," *Robotics*, vol. 10, no. 3, p. 105, Sep. 2021. [Online]. Available: https://www.mdpi.com/2218-6581/10/3/105

[31] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *J. Mach. Learn. Res.*, vol. 22, no. 30, pp. 1395–1476, 2021.

[32] X. Xiao, J. Liu, Z. Wang, Y. Zhou, Y. Qi, Q. Cheng, B. He, and S. Jiang, "Robot learning in the era of foundation models: A survey," 2023, *arXiv:2311.14379*.

[33] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large language models for robotics: A survey," 2023, *arXiv:2311.07226*.

[34] R. McCarthy, D. C. H. Tan, D. Schmidt, F. Acero, N. Herr, Y. Du, T. G. Thuruthel, and Z. Li, "Towards generalist robot learning from Internet video: A survey," 2024, *arXiv:2404.19664*.

[35] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1134–1141.

[36] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from 'in-the-wild' human videos," 2021, *arXiv:2103.16817*.

[37] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2794–2802.

[38] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

[39] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1483–1492.

[40] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6612–6619.

[41] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deep stereo: Learning to predict new views from the world's imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5515–5524.

[42] R. Garg, B. G. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. 14th Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 740–756.

[43] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[44] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6964–6974.

[45] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," 2022, *arXiv:2203.06173*.

[46] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Proc. Conf. Robot Learn.*, 2022, pp. 416–426.

[47] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," 2022, *arXiv:2203.12601*.

[48] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at Internet scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9866–9875.

[49] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning visual affordance grounding from demonstration videos," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16857–16871, Nov. 2024.

[50] H. S. Koppula and A. Saxena, "Physically grounded spatio-temporal object affordances," in *Proc. 13th Eur. Conf. Comput. Vis.-ECCV*, 2014, pp. 831–847.

[51] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, Jul. 2013.

[52] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, "Demo2Vec: Reasoning object affordances from online videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2139–2147.

[53] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13778–13790.

[54] T.-T. Do, A. Nguyen, and I. Reid, "AffordanceNet: An end-to-end deep learning approach for object affordance detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5882–5889.

[55] X. Williams and N. R. Mahapatra, "Analysis of affordance detection methods for real-world robotic manipulation," in *Proc. 9th Int. Symp. Embedded Comput. Syst. Design (ISED)*, Dec. 2019, pp. 1–5.

[56] D. I. Kim and G. S. Sukhatme, "Semantic labeling of 3D point clouds with object affordance for robot manipulation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 5578–5584.

[57] J. Ji, R. Desai, and J. C. Niebles, "Detecting human-object relationships in videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8086–8096.

[58] M. Goyal, S. Modi, R. Goyal, and S. Gupta, "Human hands as probes for interactive object understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3293–3303.

[59] T. Kwon, B. Tekin, J. Stuhmer, F. Bogo, and M. Pollefeys, "H2O: Two hands manipulating objects for first person interaction recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10118–10128.

[60] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11087–11096.

[61] B. Tekin, F. Bogo, and M. Pollefeys, "H+O: Unified egocentric recognition of 3D hand-object poses and interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4506–4515.

[62] D. Shan, R. E. L. Higgins, and D. F. Fouhey, "COHESIV: Contrastive object and hand embedding segmentation in video," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 5898–5909.

[63] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 409–419.

[64] M. Cai, K. M. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Proc. Robot., Sci. Syst.*, 2016, pp. 1–8.

[65] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, Jan. 2016.

[66] A. Pieropan, C. H. Ek, and H. Kjellstrom, "Functional object descriptors for human activity modeling," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 1282–1289.

[67] M. Ma, N. Marturi, Y. Li, A. Leonardis, and R. Stolkin, "Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos," *Pattern Recognit.*, vol. 76, pp. 506–521, Apr. 2018.

[68] J. Xin, L. Wang, K. Xu, C. Yang, and B. Yin, "Learning interaction regions and motion trajectories simultaneously from egocentric demonstration videos," *IEEE Robot. Autom. Lett.*, vol. 8, no. 10, pp. 6635–6642, Oct. 2023.

[69] F. Sener and A. Yao, "Unsupervised learning and segmentation of complex activities from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8368–8376.

[70] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[71] K. Shaw, A. Agarwal, and D. Pathak, "LEAP hand: Low-cost, efficient, and anthropomorphic hand for robot learning," 2023, *arXiv:2309.06440*.

[72] P. Mandikal and K. Grauman, "DexVIP: Learning dexterous grasping with human hand pose priors from video," in *Proc. Conf. Robot Learn.*, 2022, pp. 651–661.

[73] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: Fast monocular 3D hand and body motion capture by regression and integration," 2020, *arXiv:2008.08324*.

[74] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10967–10977.

[75] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," 2022, *arXiv:2201.02610*.

[76] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, pp. 851–866, Oct. 2023.

[77] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5026–5035.

[78] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "MERLOT: Multimodal neural script knowledge models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 23634–23651.

[79] J. C. Stroud, Z. Lu, C. Sun, J. Deng, R. Sukthankar, C. Schmid, and D. A. Ross, "Learning video representations from textual Web supervision," 2020, *arXiv:2007.14937*.

[80] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.

[81] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1708–1718.

[82] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, C. He, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "InternVid: A large-scale video-text dataset for multimodal understanding and generation," 2023, *arXiv:2307.06942*.

[83] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5842–5850.

[84] K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18973–18990.

[85] K. Grauman et al., "Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 19383–19400.

[86] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100," *Int. J. Comput. Vis.*, vol. 130, no. 1, pp. 33–55, Jan. 2022.

[87] P. Sermanet et al., "RoboVQA: Multimodal long-horizon reasoning for robotics," 2023, *arXiv:2311.00899*.

[88] A. Khazatsky et al., "DROID: A large-scale in-the-wild robot manipulation dataset," 2024, *arXiv:2403.12945*.

[89] T. Zhang, D. Li, Y. Li, Z. Zeng, L. Zhao, L. Sun, Y. Chen, X. Wei, Y. Zhan, L. Li, and X. He, "Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks," 2024, *arXiv:2405.18860*.

[90] *ActionNet: A Dataset for Dexterous Bimanual Manipulation*, Y. M. Fourier ActionNet Team, 2025.

[91] S. Jiang, H. Li, R. Ren, Y. Zhou, Z. Wang, and B. He, "Kaiwu: A multimodal manipulation dataset and framework for robot learning and human–robot interaction," 2025, *arXiv:2503.05231*.

[92] H. Zhao, X. Liu, M. Xu, Y. Hao, W. Chen, and X. Han, "TASTE-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 27683–27693.

[93] E. Perez, F. Strub, H. D. Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–11.

[94] X. Chen et al., "PaLI-X: On scaling up a multilingual vision and language model," 2023, *arXiv:2305.18565*.

[95] D. Driess et al., "PaLM-E: An embodied multimodal language model," 2023, *arXiv:2303.03378*.

[96] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," 2024, *arXiv:2406.09246*.

[97] D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. Liang Tan, L. Yunliang Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," 2024, *arXiv:2405.12213*.

[98] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, X. Wang, B. Liu, J. Fu, J. Bao, D. Chen, Y. Shi, J. Yang, and B. Guo, "CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," 2024, *arXiv:2411.19650*.

[99] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: A diffusion foundation model for bimanual manipulation," 2024, *arXiv:2410.07864*.

[100] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, H. Zhang, and M. Zhu, "GR-2: A generative video-language-action model with Web-scale knowledge for robot manipulation," 2024, *arXiv:2410.06158*.

[101] J. Bjorck et al., "GR00T N1: An open foundation model for generalist humanoid robots," 2025, *arXiv:2503.14734*.

[102] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, and X. Li, "SpatialVLA: Exploring spatial representations for visual-language-action model," 2025, *arXiv:2501.15830*.

[103] K. Black et al., "π_0: A vision-language-action flow model for general robot control," 2024, *arXiv:2410.24164*.

[104] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11975–11986.

[105] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[106] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[107] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, T.-Y. Lin, G. Wetzstein, M.-Y. Liu, and D. Xiang, "CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 1702–1713.

[108] S. Abeyruwan et al., "Gemini robotics: Bringing AI into the physical world," 2025, *arXiv:2503.20020*.

[109] J. Lee and M. S. Ryoo, "Learning robot activities from first-person human videos using convolutional future regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 472–473.

[110] H. Zhang, P.-J. Lai, S. Paul, S. Kothawade, and S. Nikolaidis, "Learning collaborative action plans from YouTube videos," in *Proc. Int. Symp. Robot. Res.*, 2022, pp. 208–223.

[111] Q. Zhang, J. Chen, D. Liang, H. Liu, X. Zhou, Z. Ye, and W. Liu, "An object attribute guided framework for robot learning manipulations from human demonstration videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6113–6119.

[112] R. Mendonca, S. Bahl, and D. Pathak, "Structured world models from human videos," 2023, *arXiv:2308.10901*.

[113] Y. Yang, Y. Li, C. Fermüller, and Y. Aloimonos, "Robot learning manipulation action plans by 'watching' unconstrained videos from the world wide Web," in *Proc. AAAI Conf. Artif. Intell.*, 2015, vol. 29, no. 1, pp. 1–7.

[114] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Daniilidis, S. Levine, and C. Finn, "Learning predictive models from observation and interaction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 708–725.

[115] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," 2023, *arXiv:2310.08576*.

[116] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," 2022, *arXiv:2207.09450*.

[117] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on YouTube," in *Proc. Robot., Sci. Syst.*, Jun. 2022, pp. 1–22.

[118] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine, "SFV: Reinforcement learning of physical skills from videos," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, Dec. 2018.

[119] X. Deng, J. Liu, H. Gong, H. Gong, and J. Huang, "A human–robot collaboration method using a pose estimation network for robot learning of assembly manipulation trajectories from demonstration videos," *IEEE Trans. Ind. Informat.*, vol. 19, no. 5, pp. 7160–7168, May 2023.

[120] Y.-T.-A. Sun, H.-C. Lin, P.-Y. Wu, and J.-T. Huang, "Learning by watching via keypoint extraction and imitation learning," *Machines*, vol. 10, no. 11, p. 1049, Nov. 2022.

[121] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 7827–7834.

[122] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[123] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–12.

[124] M.-Y. Liu, T. M. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–5.

[125] T. I. Erdei, T. P. Kapusi, A. Hajdu, and G. Husi, "Image-to-image translation-based deep learning application for object identification in industrial robot systems," *Robotics*, vol. 13, no. 6, p. 88, Jun. 2024.

[126] P. Sharma, D. Pathak, and A. Gupta, "Third-person visual imitation learning via decoupled hierarchical controller," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–8.

[127] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[128] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine, "AVID: Learning multi-stage tasks via pixel-level translation of human videos," 2019, *arXiv:1912.04443*.

[129] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[130] S. Dasari and A. Gupta, "Transformers for one-shot visual imitation," in *Proc. Conf. Robot Learn.*, 2020, pp. 2071–2084.

[131] J. Li, T. Lu, X. Cao, Y. Cai, and S. Wang, "Meta-imitation learning by watching video demonstrations," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–9. [Online]. Available: https://openreview.net/forum?id=KTPuIsx4pmo

[132] J. Liu, L. He, Y. Kang, Z. Zhuang, D. Wang, and H. Xu, "CEIL: Generalized contextual imitation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 75491–75516.

[133] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1118–1125.

[134] S. Yang, W. Zhang, W. Lu, H. Wang, and Y. Li, "Cross-context visual imitation learning from demonstrations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 5467–5473.

[135] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[136] V. Petrík, M. Tapaswi, I. Laptev, and J. Šivic, "Learning object manipulation skills via approximate state estimation from real videos," in *Proc. Conf. Robot Learn.*, 2020, pp. 296–312.

[137] K. Zorina, J. Carpentier, J. Sivic, and V. Petrík, "Learning to manipulate tools by aligning simulation to video demonstration," *IEEE Robot. Autom. Lett.*, vol. 7, no. 1, pp. 438–445, Jan. 2022.

[138] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn, "Reinforcement learning with videos: Combining offline observations with interaction," in *Proc. Conf. Robot Learn.*, 2020, pp. 339–354.

[139] D. Xu, S. Nair, Y. Zhu, J. Gao, A. Garg, L. Fei-Fei, and S. Savarese, "Neural task programming: Learning to generalize across hierarchical tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3795–3802.

[140] O. Mees, M. Merklinger, G. Kalweit, and W. Burgard, "Adversarial skill networks: Unsupervised robot skill learning from video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4188–4194.

[141] E. Chane-Sane, C. Schmid, and I. Laptev, "Learning video-conditioned policies for unseen manipulation tasks," 2023, *arXiv:2305.06289*.

[142] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[143] A. Bonardi, S. James, and A. J. Davison, "Learning one-shot imitation from humans without humans," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3533–3539, Apr. 2020.

[144] P. Sharma, L. Mohan, L. Pinto, and A. Gupta, "Multiple interactions made easy (MIME): Large scale demonstrations data for imitation," in *Proc. Conf. robot Learn.*, 2018, pp. 906–915.

[145] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[146] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[147] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," 2018, *arXiv:1807.06158*.

[148] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Comput. Surveys*, vol. 55, no. 13s, pp. 1–40, Dec. 2023.

[149] D. Pathak, P. Mahmoudieh, G. Luo, P. Agrawal, D. Chen, F. Shentu, E. Shelhamer, J. Malik, A. A. Efros, and T. Darrell, "Zero-shot visual imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2131–21313.

[150] Z. Mandi, F. Liu, K. Lee, and P. Abbeel, "Towards more generalizable one-shot visual imitation learning," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2434–2444.

[151] W. Goo and S. Niekum, "One-shot learning of multi-step tasks from observation via activity localization in auxiliary video," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7755–7761.

[152] M. J. Kim, J. Wu, and C. Finn, "Giving robots a hand: Learning generalizable manipulation with eye-in-hand human video demonstrations," 2023, *arXiv:2307.05959*.

[153] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[154] A. Zhou, E. Jang, D. Kappler, A. Herzog, M. Khansari, P. Wohlhart, Y. Bai, M. Kalakrishnan, S. Levine, and C. Finn, "Watch, try, learn: Meta-learning from demonstrations and reward," 2019, *arXiv:1906.03352*.

[155] M. Sieb, X. Zhou, A. Huang, O. Kroemer, and K. Fragkiadaki, "Graph-structured visual imitation," in *Proc. Conf. Robot Learn.*, 2019, pp. 979–989.

[156] K. Ramachandruni, M. Babu, A. Majumder, S. Dutta, and S. Kumar, "Attentive task-net: Self supervised task-attention network for imitation learning using video demonstration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4760–4766.

[157] G. Garcia-Hernando, E. Johns, and T.-K. Kim, "Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 9561–9568.

[158] K. Pertsch, R. Desai, V. Kumar, F. Meier, J. J. Lim, D. Batra, and A. Rai, "Cross-domain transfer via semantic skill imitation," 2022, *arXiv:2212.07407*.

[159] Y. Aytar, T. Pfaff, D. Budden, T. Paine, Z. Wang, and N. D. Freitas, "Playing hard exploration games by watching YouTube," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 2930–2941.

[160] B. C. Stadie, P. Abbeel, and I. Sutskever, "Third-person imitation learning," 2017, *arXiv:1703.01703*.

[161] S. Huo, A. Duan, L. Han, L. Hu, H. Wang, and D. Navarro-Alarcon, "Efficient robot skill learning with imitation from a single video for contact-rich fabric manipulation," 2023, *arXiv:2304.11801*.

[162] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[163] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg, "Concept2Robot: Learning manipulation concepts from instructions and human demonstrations," *Int. J. Robot. Res.*, vol. 40, nos. 12–14, pp. 1419–1434, 2021.

[164] C. Jiang, M. Dehghan, and M. Jagersand, "Understanding contexts inside robot and human manipulation tasks through vision-language model and ontology system in video streams," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 8366–8372.

[165] S. Yang, W. Zhang, W. Lu, H. Wang, and Y. Li, "Learning actions from human demonstration video for robotic manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 1805–1811.

[166] S. Yang, W. Zhang, R. Song, J. Cheng, H. Wang, and Y. Li, "Watch and act: Learning robotic manipulation from visual demonstration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 7, pp. 4404–4416, Jul. 2023.

[167] H. Zhang, J. Zhong, and S. Nikolaidis, "Zero-shot imitating collaborative manipulation plans from YouTube cooking videos," 2019, *arXiv:1911.10686*.

[168] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis, "Translating videos to commands for robotic manipulation with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3782–3788.

[169] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, I. Gilitschenski, Y. Bisk, and D. Dwibedi, "Vid2Robot: End-to-end video-conditioned policy learning with cross-attention transformers," 2024, *arXiv:2403.12943*.

[170] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," 2024, *arXiv:2401.11439*.

[171] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," 2023, *arXiv:2302.02011*.

[172] G. Thomas, C.-A. Cheng, R. Loynd, F. V. Frujeri, V. Vineet, M. Jalobeanu, and A. Kolobov, "PLEX: Making the most of the available data for robotic manipulation pretraining," 2023, *arXiv:2303.08789*.

[173] O. Mees, L. Hermann, and W. Burgard, "What matters in language conditioned robotic imitation learning over unstructured data," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11205–11212, Oct. 2022.

[174] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7327–7334, Jul. 2022.

[175] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "VIMA: General robot manipulation with multimodal prompts," 2022, *arXiv:2210.03094*.

[176] T. Cui, T. Zhou, Z. Peng, M. Hu, H. Lu, H. Li, G. Chen, M. Wang, and Y. Yue, "Human demonstrations are generalizable knowledge for robots," 2023, *arXiv:2312.02419*.

[177] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," 2023, *arXiv:2312.13139*.

[178] H. He, C. Bai, L. Pan, W. Zhang, B. Zhao, and X. Li, "Learning an actionable discrete diffusion policy via large-scale actionless video pre-training," 2024, *arXiv:2402.14407*.

[179] NVIDIA. *ISAAC Sim*. [Online]. Available: https://github.com/isaac-sim/IsaacSim

[180] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 5026–5033.

[181] E. Coumans and Y. Bai. (2021). *Pybullet, a Python Module for Physics Simulation for Games, Robotics and Machine Learning*. [Online]. Available: http://pybullet.org

[182] O. Contributors. (May 11, 2010). *Open Source Computer Vision Library (OpenCV)*. Accessed: Aug. 9, 2025. [Online]. Available: https://github.com/opencv/opencv

[183] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for perceiving and processing reality," in *Proc. 3rd Workshop Comput. Vis. AR/VR IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–4. [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitleMay1MediaPipeCVPRCV4ARVRWorkshop2019.pdf

[184] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[185] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The EPIC-KITCHENS dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 720–736.

[186] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2248–2255.

[187] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[188] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.
[189] C. Dima, M. Hebert, and A. Stentz, "Enabling learning from large datasets: Applying active learning to mobile robotics," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 1, Apr. 2004, pp. 108–114.
[190] B. Settles, "Active learning literature survey," Tech. Rep., 2009.
[191] M. Wulfmeier, A. Bewley, and I. Posner, "Addressing appearance change in outdoor robotics with adversarial domain adaptation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 1551–1558.
[192] A. Paudel, "Learning for robot decision making under distribution shift: A survey," 2022, *arXiv:2203.07558*.
[193] W. Huang, I. Mordatch, P. Abbeel, and D. Pathak, "Generalization in dexterous manipulation via geometry-aware multi-task learning," 2021, *arXiv:2111.03062*.
[194] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3758–3765.
[195] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, "Learning modular neural network policies for multi-task and multi-robot transfer," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2169–2176.
[196] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8943–8950.
[197] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 5824–5836.
[198] R. Wolf, Y. Shi, S. Liu, and R. Rayyes, "Diffusion models for robotic manipulation: A survey," 2025, *arXiv:2504.08438*.
[199] S. Parab. (2024). *The Rise of Diffusion Models in Imitation Learning*. Accessed: Aug. 9, 2025. [Online]. Available: https://www.trossenrobotics.com/post/the-rise-of-diffusion-models-in-imitation-learning
[200] N. Ingelhag, J. Munkeby, J. van Haastregt, A. Varava, M. C. Welle, and D. Kragic, "A robotic skill learning system built upon diffusion policies and foundation models," in *Proc. 33rd IEEE Int. Conf. Robot Human Interact. Commun. (ROMAN)*, Aug. 2024, pp. 748–754.
[201] W. Yan, O. Watkins, S. James, R. Okumura, T. Darrell, and P. Abbeel. (2025). *Task-Specific World Models for Robotic Manipulation*. Accessed: Aug. 9, 2025. [Online]. Available: https://bcommons.berkeley.edu/task-specific-world-models-robotic-manipulation
[202] F. Zhu, H. Wu, S. Guo, Y. Liu, C. Cheang, and T. Kong. (2024). *ManiWM: An Interactive World Model for Real-Robot Manipulation*. [Online]. Available: https://openreview.net/forum?id=aVyJwS1fqQ
[203] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3019–3026, Apr. 2020.
[204] Y. Wen, J. Lin, Y. Zhu, J. Han, H. Xu, S. Zhao, and X. Liang, "VidMan: Exploiting implicit dynamics from video diffusion model for effective robot manipulation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 41051–41075.
[205] H. Xiong, H. Fu, J. Zhang, C. Bao, Q. Zhang, Y. Huang, W. Xu, A. Garg, and C. Lu, "RoboTube: Learning household manipulation from human videos with simulated twin environments," in *Proc. 6th Conf. Robot Learn.*, vol. 205, Dec. 2023, pp. 1–10. [Online]. Available: https://proceedings.mlr.press/v205/xiong23a.html
[206] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6979–6987.
[207] T. E. Lee, "Causal robot learning for manipulation," Ph.D. dissertation, Dept. Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, Jul. 2024.
[208] Y. Li, A. Torralba, A. Anandkumar, D. Fox, and A. Garg, "Causal discovery in physical systems from videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9180–9192.
[209] Y. Li, "Deep causal learning for robotic intelligence," *Frontiers Neurorobotics*, vol. 17, Feb. 2023, Art. no. 1128591.
[210] A. Méndez-Molina, E. F. Morales, and L. E. Sucar, "Causal discovery and reinforcement learning: A synergistic integration," in *Proc. 11th Int. Conf. Probabilistic Graph. Models*, vol. 186, Oct. 2022, pp. 421–432. [Online]. Available: https://proceedings.mlr.press/v186/mendez-molina22a.html
[211] L. Castri, G. Beraldo, and N. Bellotto, "Causality-enhanced decision-making for autonomous mobile robots in dynamic environments," 2025, *arXiv:2504.11901*.
[212] T. E. Lee, S. Vats, S. Girdhar, and O. Kroemer, "SCALE: Causal learning and discovery of robot manipulation skills using simulation," in *Proc. 7th Conf. Robot Learn.*, vol. 229, Nov. 2023, pp. 2229–2256. [Online]. Available: https://proceedings.mlr.press/v229/lee23b.html
[213] L. Castri, S. Mghames, and N. Bellotto, "From continual learning to causal discovery in robotics," in *Proc. 1st AAAI Bridge Program Continual Causality*, Feb. 2023, pp. 85–91. [Online]. Available: https://proceedings.mlr.press/v208/castri23a.html
[214] CVPR Workshop. (2024). *Causal and Object-Centric Representations for Robotics*. Accessed: Aug. 9, 2025. [Online]. Available: https://corrworkshop.github.io/

CHRISANTUS EZE (Graduate Student Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from the Federal University of Technology, Owerri, Nigeria. He is currently pursuing the Ph.D. degree in computer science with Oklahoma State University.

From 2019 to 2021, he was a Software Engineer with Seamfix Ltd., Lagos, Nigeria. Since January 2022, he has been the Ph.D. Student with the Computer Science Department, Oklahoma State University, Stillwater. He is the author of three conference publications. His research interests include improving long-horizon and spatial reasoning in robots for manipulation. To achieve this, he proposes and implores techniques, algorithms and systems for computer vision, foundation models, and imitation learning.

CHRISTOPHER CRICK (Member, IEEE) received the B.A. degree in history and the M.S. degree in computer science from Harvard University and the Ph.D. degree in computer Science from Yale University, in 2009.

He joined Brown University as a Postdoctoral Fellow. He joined Oklahoma State University as a Professor, in 2018, and carries out research at the intersection of robotics and explainable AI. His research agenda is motivated by two overarching, complementary goals: to ground models of developmental psychology and cognitive science in realized, embodied computational systems; and to use models of human cognition and social development to improve robot control and learning, human–robot interaction, and artificial intelligence in general.