Towards Generalist Robot Learning from Internet Video: A Survey

ROBERT MCCARTHY*, University College London, United Kingdom
DANIEL C.H. TAN, University College London, United Kingdom
DOMINIK SCHMIDT, Weco AI, United Kingdom
FERNANDO ACERO, University College London, United Kingdom
NATHAN HERR, University College London, United Kingdom
YILUN DU, Massachusetts Institute of Technology, United States of America
THOMAS G. THURUTHEL, University College London, United Kingdom
ZHIBIN LI, University College London, United Kingdom

Scaling deep learning to massive and diverse internet data has driven remarkable breakthroughs in domains such as video generation and natural language processing. Robot learning, however, has thus far failed to replicate this success and remains constrained by a scarcity of available data. Learning from Videos (LfV) methods aim to address this data bottleneck by augmenting traditional robot data with large-scale internet video. This video data provides foundational information regarding physical dynamics, behaviours, and tasks, and can be highly informative for general-purpose robots.

This survey systematically examines the emerging field of LfV. We first outline essential concepts, including detailing fundamental LfV challenges such as distribution shift and missing action labels in video data. Next, we comprehensively review current methods for extracting knowledge from large-scale internet video, overcoming LfV challenges, and improving robot learning through video-informed training. The survey concludes with a critical discussion of future opportunities. Here, we emphasize the need for scalable foundation model approaches that can leverage the full range of available internet video and enhance the learning of robot policies and dynamics models. Overall, the survey aims to inform and catalyse future LfV research, driving progress towards general-purpose robots.

JAIR Track: Surveys

JAIR Associate Editor: Nick Hawes

JAIR Reference Format:
Robert McCarthy, Daniel C.H. Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G. Thuruthel, and Zhibin Li. 2025. Towards Generalist Robot Learning from Internet Video: A Survey. *Journal of Artificial Intelligence Research* 83, Article 12 (July 2025), 48 pages. DOI: 10.1613/jair.1.17400

*Corresponding Author.

Authors' Contact Information: Robert McCarthy, robert.mccarthy.22@ucl.ac.uk, ORCID: 0000-0002-2140-6988, University College London, United Kingdom; Daniel C.H. Tan, daniel.tan.22@ucl.ac.uk, ORCID: 0000-0003-1067-8432, University College London, United Kingdom; Dominik Schmidt, ORCID: 0000-0001-8208-6087, schmidtdominik30@gmail.com, Weco AI, United Kingdom; Fernando Acero, fernando.acero@ucl.ac.uk, ORCID: 0000-0002-5470-8122, University College London, United Kingdom; Nathan Herr, nathan.herr.19@ucl.ac.uk, ORCID: 0009-0002-1915-6980, University College London, United Kingdom; Yilun Du, yilundu@mit.edu, ORCID: 0000-0001-6792-5946, Massachusetts Institute of Technology, United States of America; Thomas G. Thuruthel, t.thuruthel@ucl.ac.uk, ORCID: 0000-0003-0571-1672, University College London, United Kingdom; Zhibin Li, alex.li@ucl.ac.uk, ORCID: 0000-0002-6357-7419, University College London, United Kingdom.

DOI: 10.1613/jair.1.17400

1 Introduction

A *generalist robot* should be capable of performing a broad range of physical tasks in unstructured real-world environments. It should maintain high-level reasoning and planning abilities, along with low-level physical skills, such as dexterous manipulation. It should adapt to unseen and unexpected scenarios. It should operate in unstructured settings by perceiving the world through imperfect partial observations (e.g., visual and tactile sensing).

Such a robot would be highly useful in many practical applications (e.g., household or factory tasks). Nevertheless, obtaining generalist robots remains a grand challenge in robotics. Classical robotics techniques are insufficient as they cannot handle unseen and unstructured scenarios [137]. More recent machine learning (i.e., robot learning) techniques are more promising [206, 113]. These approaches have achieved skillful robot control in narrow settings [336, 127], but generalization to unseen settings remains a challenge.

Robot learning can take inspiration from advances in other domains. Deep learning has recently provided remarkable improvements in natural language processing [193], image generation [24], and video generation [38]. This has been achieved by training expressive architectures [274] on massive, diverse datasets scraped from the internet. Here, 'scaling laws' have shown performances to consistently and predictably improve with increased computational power and data [123]. The use of diverse internet data has facilitated a transition from task-specific models to monolithic 'foundation models' with more general capabilities [193].

There is evidence that these deep learning techniques [36, 264] and scaling laws [97, 222, 40] can transfer to robotics and control. This offers a path towards more general-purpose robot capabilities. However, obtaining large-scale data is a challenge in robotics. Robotics faces a *chicken-and-egg* problem: data cannot be easily collected due to limited robot capabilities (deploying limited robots to collect real-world data can be ineffective and dangerous), and capabilities cannot easily be improved due to the lack of data (see Figure 1).

How can we overcome this data bottleneck? One possibility is to use humans to collect robot data [36, 129]. However, this is expensive and can be difficult in tasks requiring skill. Another option is to leverage simulation [336, 127]. However, simulation comes with issues related to flawed simulated physics and difficulties creating and training on a suitable diversity of simulated environments and tasks. A final option is, like previous deep learning successes [193, 24], to leverage the vast quantities of data available on the internet.

While any practical approach may leverage all complementary sources of data, in this survey we focus on learning from internet data. Specifically, our interest lies in *internet video data*. Our reasoning here is threefold:

(1) *Relevant information content.* In contrast to internet-scraped text or image data, internet video can uniquely offer information regarding the physics and dynamics of the world, and information regarding human behaviours and actions [308]. Crucially, internet video has coverage over many behaviours and tasks relevant to a general-purpose robot (e.g., household chores).
(2) *High in quantity and diversity.* There are huge quantities of video data freely available on the internet [249]. Importantly, this data is highly diverse. The largest open-source robot dataset [195] pales in comparison, both in terms of quantity and diversity of the data (see Figure 8).
(3) *Internet video is relatively untapped.* The use of real and simulated robot data has been extensively explored [36, 264, 8, 127]. Meanwhile, leveraging pretrained text and image foundation models has become increasingly common in robotics [6, 157, 37, 238]. However, the use of internet video data in robotics is in more nascent stages.

By leveraging large, diverse internet video data during robot learning, we hope to obtain a number of benefits (Section 3.3). This includes obtaining improved generalization beyond the available robot data, versus approaches that rely solely on small, narrow robot datasets (see Figure 2). Indeed, recent progress in the emerging field of Learning from Videos (LfV) (Section 3.2) has been promising, demonstrating evidence of this. This has

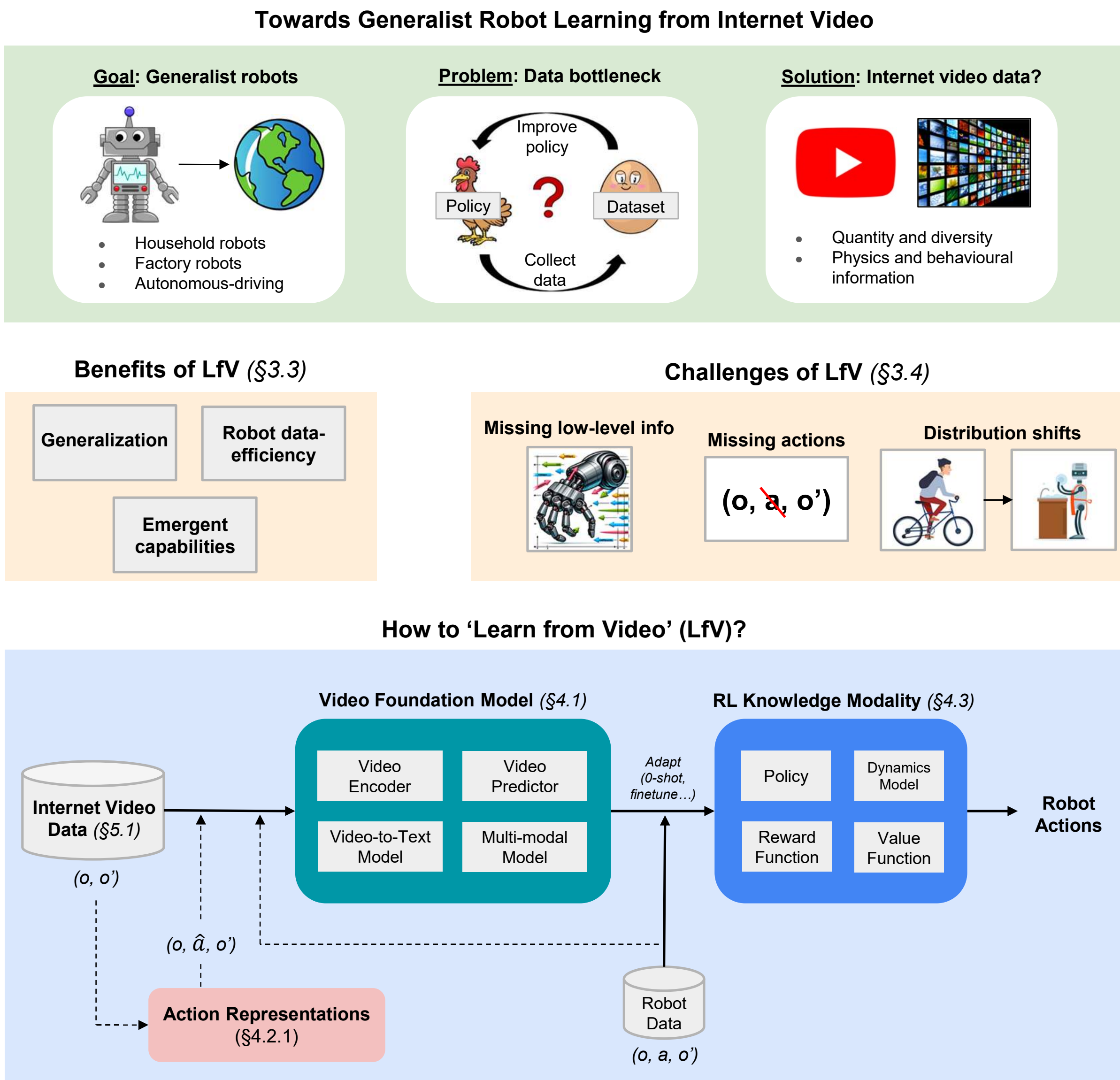


Fig. 1. **An overview of the key concepts and taxonomies in this survey.** The top green box presents the high-level motivation behind LfV. The middle orange boxes highlight the benefits (Section 3.3) and challenges (Section 3.4) of LfV. The bottom blue box visualises possible components in a pipeline for learning from large-scale internet video, as per the taxonomies presented in the survey. Large internet video datasets (Section 5.1) can be used to pretrain (video) foundation models (Section 4.1). These models can be adapted (e.g., via zero-shot transfer or finetuning) into reinforcement learning (RL) 'knowledge modalities' [291] for use in the robot domain (Section 4.3). The diagram additionally highlights that action representations (Section 4.2.1) can be used to mitigate the issue of missing action labels in video.

included work leveraging large-scale video prediction models to act as robot dynamics models [306, 40], and work leveraging both robot data and internet video to train foundation models for robotics [251].

Nevertheless, leveraging internet video is non-trivial and comes with a number of challenges (Section 3.4). First, video is a challenging data modality: it is high-dimensional, noisy, and poorly labelled. Second, video lacks information critical to robotics, including action labels and low-level force and proprioceptive information. Moreover, there may be various distribution shifts between internet video and the robot domain. Given these challenges, two key questions for LfV research are: (i) 'How to extract relevant knowledge from internet video?' and (ii) 'How to apply video-extracted knowledge to robotics?'.

In this survey, we review literature that attempts to answer these questions [1]. For the first question, we survey video foundation model techniques promising for extracting knowledge from large-scale, heterogenous internet video (Section 4.1). These techniques may serve as the backbone of scalable LfV approaches. For the second question, we perform a thorough analysis of literature that explicitly addresses LfV challenges (Section 4.2), before reviewing methods that directly utilise video data to aid robot learning (Section 4.3). We conclude the survey by discussing challenges and opportunities for future LfV research (Section 6). Here, we recommend focusing on scalable approaches that can leverage the full range of available internet video data, and focusing on learning policies and dynamics models to best obtain the benefits of LfV.

Existing LfV research has shown promising signs of life, and this is a field ripe for further progress. Of course, it is unclear whether scaling deep learning will take us all the way to general-purpose robots (Section 6.4), and there are other research directions that should be pursued in parallel (Section 2). Nevertheless, with sufficient effort and innovation, it seems likely that scaling—with the help of internet video data—can provide significant gains in the near-future. We hope this comprehensive survey can encourage and inform these future LfV research efforts, serving to facilitate our progress towards the creation of generalist robots.

2 Background

This section provides further details on prior literature in deep learning and robot learning. This serves to add context regarding where LfV approaches (and our survey) fit into the broader research landscape.

Scaling Deep Learning. Empirically-derived 'scaling laws' have shown deep learning performances to consistently and predictably improve with increased computation, model parameter count, and data [123]. Recent successes in deep learning [39, 193, 70, 24] have largely been driven by such scaling, aided by the use of improved model architectures – transformer [274] and diffusion models [99], and the use of self-supervised objectives that allow for learning from unlabelled internet data. The term 'foundation model' is commonly used to refer to a large, general-purpose model, pretrained on large diverse data, that can solve wide varieties of downstream tasks [34].

Large language models (LLMs) are transformer models with billions of parameters [70], pretrained via a simple next-token prediction objective on huge text datasets scraped from the internet [39, 189]. This pretraining scheme allows the model to learn a 'foundation' of declarative and procedural knowledge. The pretrained LLM can be finetuned into a more functional product via supervised learning on high-quality curated data and/or via reinforcement learning from human (or AI) feedback (RLHF) [194, 16]. Meanwhile, image generation has also benefited from the use of large-scale internet data and from diffusion architectures [218, 24]. Multi-modal vision-language models (VLMs), trained on large-scale internet data to take both images and text as input, have progressed visual understanding [9, 193, 263]. We detail *video* foundation model efforts in Section 4.1.

Scaling Robot Learning. There is a hope that scaling robot learning, in a similar fashion to other domains, may improve the generality of our robots [195]. However, there are well-known challenges here. Unlike in video

[1]See Figure 1 for more details regarding the structure of the survey.

games [229], scaling online reinforcement learning (RL) is not a practical solution for real-world robotics: it is time-consuming, costly, and dangerous [260]. Meanwhile, offline learning solutions [146, 36] lack suitably large-scale, high-quality robotic datasets [5]. A number of approaches have been proposed to address these challenges and scale up robotic datasets:

- *Simulated data.* The use of simulation is promising for scaling up online learning while avoiding avoiding difficulties associated with the real-world [336, 127, 190]. However, the use of simulation presents a number of issues [25]. (1) Inaccuracies in low-level physics create a ‘sim-to-real’ gap [329] that must be overcome. A partial solution here is to employ domain randomization [266]. (2) Manually creating a suitable diversity of simulated environments and tasks for generalist robotics is a challenge. Recent works seek to tackle this using procedural environment generation [62, 262] and foundation-model-assisted environment design [292, 76, 286]. (3) We often lack a policy to collect high-quality data in the simulation. Solutions here have included the use of an automated curriculum during RL training [200, 158] and collecting data using: specialist policies [259], models with access to privileged simulation information [92, 276, 57], or bootstrapping data collection from human demonstrations [179, 177].
- *Real world data.* Data collected via human teleoperation [36, 129] and pooled data from multiple academic labs [195] have been used to train initial robot foundation models [36, 264]. Other works have investigated methods for automating data collection to improve scalability [35, 5, 304]. Several companies have demonstrated evidence of infrastructure suitable for large-scale robot data-collection [251, 115]. However, the largest open-source robot dataset [195] is still significantly smaller and less diverse than internet-scale data (see Figure 8).
- *Internet data.* Robot learning has benefited from the use of broad internet data. Image and video data have been used to pretrain visual representations for robotics [277, 188]. Foundational VLMs and LLMs have been used to help define reward functions for the robot learner [258, 69, 317, 133]. LLMs have been employed as high-level planners in long-horizons tasks [6, 112]. Some works have finetuned a standard internet-pretrained foundation model into a robot foundation model [37, 131]. Others have jointly pretrained on both internet data and action-labelled robot data [217, 251]. We elaborate on how internet *video* data has been used to aid robot learning in Section 4.3.

Other Paths to Generalist Robots. Scaling end-to-end deep learning approaches that train a monolithic ‘foundation’ model is one potential path to general-purpose robot capabilities. Other paths could involve combining learning with the use of stronger inductive biases, more structure, or classical robotics techniques. These may help address robotics-specific challenges and potential limitations of scaling end-to-end deep learning (Kumar [139], Section 6.3). Modular approaches have been proposed which use specialist models for different skills or tasks [303, 145], or use a hierarchy to separate out high-level planning from low-level control [117, 144, 320]. Learning has been combined with explicit planning algorithms [182, 111] and environment representation techniques [48, 178, 201]. Structure related to the robot embodiment [282, 236] or the physics of the system [167, 147] can be leveraged, while some approaches carefully choose and process their observation and action spaces [299, 54]. As these ‘alternative’ approaches and deep learning are often combined, they can be complementary: advances in scaling deep learning can benefit alternative approaches, and vice versa (e.g., alternative approaches can collect data for monolithic deep learning models; Dalal et al. [57]).

Related Surveys. Ravichandar et al. [216] and Gavenski et al. [79] survey imitation learning methods; Prudencio et al. [207] review offline RL methods; while Wulfmeier et al. [291] highlight the promise of transferring knowledge from a source to a target domain in RL. Yang et al. [307] and Hu et al. [107] review the use of foundation models in decision-making and robotics. Relevant works in the video ML literature have included reviews of self-supervised video learning [223], foundation models for video understanding [173], and LLMs for video understanding [261, 324].

Torabi et al. [272] review imitation learning from observational data, while Yang et al. [308] advocate for the use of video (and video generation) as a unified interface to absorb internet knowledge and represent diverse tasks. Eze and Crick [75] review video-based learning approaches for robot manipulation. In contrast to this work, our survey places a stronger focus on approaches that: (i) can scale to large, diverse internet video data, and (ii) can yield general-purpose robotic capabilities.

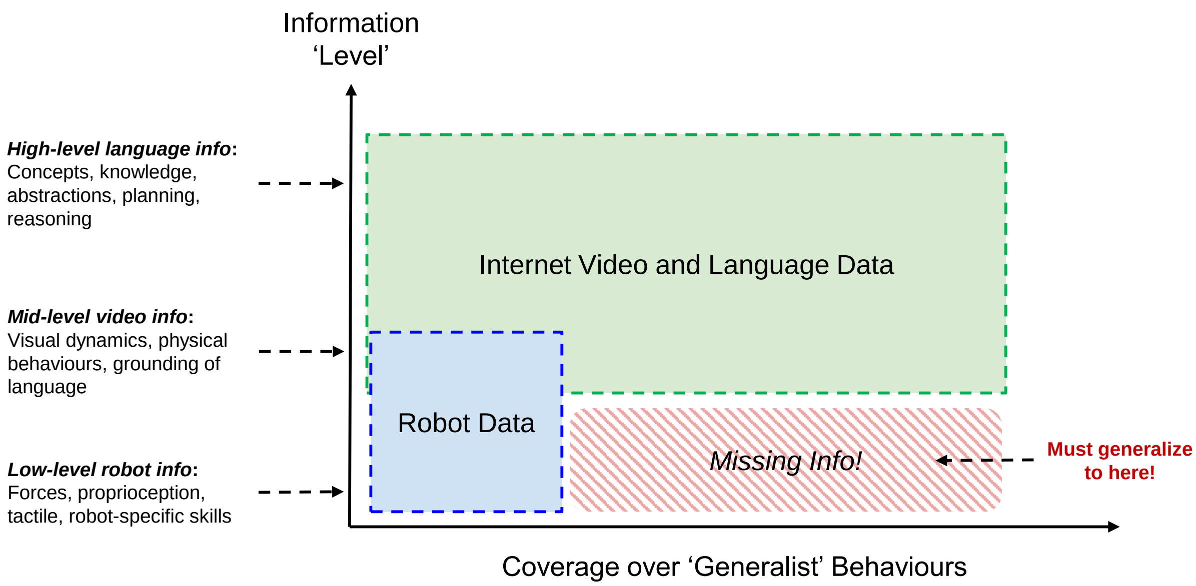


Fig. 2. **Generalization in the Learning from Videos (LfV) setting.** The x-axis indicates the range of behaviours expected from a generalist robot. The y-axis indicates the 'levels' of information contained in data. The figure demonstrates that internet data has better coverage over desired behaviours than narrow robot datasets, but lacks crucial low-level information essential to robotics. Generalising beyond the robot data despite this missing low-level information is a key LfV challenge. See Sections 3.3 and 3.4 for further discussion.

3 Preliminaries

In this section, we introduce useful preliminary concepts related to LfV. First, the reinforcement learning (Section 3.1) and LfV (Section 3.2) settings are formalized. Next, we perform a deeper dive into the potential benefits of LfV (Section 3.3) and the fundamental challenges that come with learning from videos in robotics (Section 3.4).

3.1 Reinforcement Learning (RL)

Formalism. In reinforcement learning (RL), an agent observes its environment, takes an action, and receives a reward after the state of the environment changes. This can be formalised as a Markov Decision Process (MDP) consisting of the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the transition probability $p(s_{t+1}|s_t, a_t)$ of reaching state $s_t + 1$ from state s_t when executing action a_t. The agent's behaviour is given in terms of a policy $\pi(a_t|s_t)$. When aspects of the state cannot be observed, the environment is termed a Partially Observable MDP (POMDP) and the agent only has access to observations $o_t \in O$ that are partial mappings of the state $o_t = f(s_t)$. Unless stated otherwise, we will use always a POMDP setting.

The agent aims to maximise its sum of discounted future rewards, commonly referred to as its 'returns'. This is captured by the objective in Equation 1:

$$J(\pi) = \mathbb{E}_{[p(s_0),p(s_{t+1}|s_t,a_t),\pi(a_t|s_t)]_{t=0...T}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{1}$$

where γ is the discount factor, $r_t = r(s_t, a_t)$ is the reward, and $\rho(s_0)$ is the initial state distribution. The optimal policy $\pi^*(a_t|s_t)$ maximises Equation 1.

In the LfV setting, the POMDP is some real-world environment and $\pi(a_t|s_t)$ controls a robot within this environment.

Knowledge Modalities. An RL Knowledge Modality (KM) is some function learned from data that represents specific types of RL-related knowledge (see the work of Wulfmeier et al. [291] for more details). We make use of the notion of a KM throughout our paper, including to help define our main taxonomy of LfV methods in Section 4.3. We will commonly refer to the following KMs:

- A *policy,* $\pi(a_t|s_t)$ is a mapping from states to actions directly describing the agent's behaviour.
- A *state or state-action value function,* $V_\pi(s_t)$ *or* $Q_\pi(s_t, a_t)$ maps from states or state-action pairs to the expected future return when acting under a particular policy π.
- A *dynamics model,* $p(s_{t+1}|s_t, a_t)$ predicts the next state given the current state and action.
- A *reward model,* $r(s_t, a_t)$ predicts the reward provided by the environment for a specific state-action pair.

Transfer Mechanisms. RL KMs can be pretrained on a source MDP or dataset and can then be subsequently transferred to a target MDP [291]. In the case of LfV, this involves learning a KM from a video dataset and adapting it to the robot MDP (usually via the use of robot data). We will commonly refer to the following transfer mechanisms in our paper (see the work of Wulfmeier et al. [291] for more details): *Generalisation and zero-shot transfer* involves using a pretrained KM directly in the target setting, without further fine-tuning or adaptation [74]. We use *Fine-tuning* and *Representation transfer* interchangeably in this survey to refer to methods that train a model in the robot domain that is partially composed of parts of a pretrained KM [188, 174]. *Hierarchy: conditioning* involves using a pretrained KM to condition a new KM being trained in the target MDP [227, 278, 288].

3.2 Learning from Videos (LfV): Formalism and Assumptions

Formalism. We assume an LfV method has access to a video dataset D_{video}. We denote a video clip as $\tau = (o_0, o_1, ..., o_T)$, where τ is the full clip and each o is an RGB image observation. Optionally, D_{video} may come paired with language annotations or annotations of other modalities. We also assume access to a robot dataset D_{robot}. This contains trajectories of transition tuples (o_t, a_t, r_t, o_{t+1}), where r_t may be missing, and where each o contains an image observation and may also contain other information (e.g., tactile information). The goal of LfV is to leverage our combined LfV dataset, $D_{\text{LfV}} = \{D_{\text{robot}}, D_{\text{video}}\}$, to obtain an improved $\pi(a_t|s_t)$ versus when learning from D_{robot} alone.

LfV in the Generalist Robot Setting. Though D_{video} may come from any source, in this survey we are interested in methods that can leverage large-scale video data gathered from the internet. We will generally assume that such a dataset consists primarily of videos of humans and has broad coverage over many common physical tasks that humans perform. We assume the robot must perform tasks that, to some extent, seen in D_{video}.

We loosely define a ‘generalist’/general-purpose robot as one that can perform a diverse range of every-day physical human tasks in unstructured real-world settings. Such settings are POMDPs, where the robot must rely heavily on visual observations. Throughout the survey, unless stated otherwise, we assume the general-purpose robot has an embodiment and affordances similar to those of a human. We subsequently assume the robot can execute tasks in a physically similar manner to how a human would.

Under the above assumptions, internet video can be particularly informative to the robot: it provides extensive information regarding how relevant embodiments can perform relevant tasks and behaviours.

Limitations of these assumptions. First, for certain robot tasks, embodiments very different from that of a human may prove more effective — e.g., aerial drones [10] or quadrupedal robots [28, 303] that can better traverse treacherous terrains. Second, D_{video} may not have good coverage over all the tasks we want the robot to perform: some important actions and tasks may be underrepresented in the video dataset — e.g., specialized industrial tasks [273]. Additionally, the robot may need to perform tasks not commonly performed by humans — e.g., planetary exploration [224]. Nevertheless, even in these cases, internet video can still provide generally information regarding the world and physical behaviour.

3.3 Benefits of LfV

Robotic datasets are expensive to acquire and thus are currently task-specific or relatively narrow [195]. In contrast, diverse video data is freely available in vast quantities on the internet. In this section, we outline the specific benefits we hope to obtain from methods that leverage this video data.

Generalization beyond D_{robot}. LfV offers the exciting possibility of improved generalization beyond narrow robot datasets, D_{robot}, to the full space of tasks covered in the video dataset, D_{video} [75]. An argument for why this could be the case is as follows. First, consider a D_{robot} that has good coverage over the low-level skills or ‘atomic’ actions required from the robot (e.g., specific grasping motions or locomotion skills) [53]. Now consider a task unseen in D_{robot} but seen in D_{video}. Our combined LfV dataset ($D_{\text{LfV}} = \{D_{\text{robot}}, D_{\text{video}}\}$) may contain most information required to complete the task. D_{robot} provides information regarding how to execute the low-level robot skills, whilst the human actions in D_{video} provides higher-level information regarding how to complete the overall task (e.g., visual information regarding required movements and steps). Thus, a suitable LfV method may be capable of leveraging D_{video} to generalise beyond D_{robot} and solve the task. There is preliminary evidence of such generalization in LfV-related literature [37, 67, 289, 278]. Figure 2 explores this generalization setting in more detail.

Emergent Capabilities. Learning from internet-scale video may yield capabilities qualitatively beyond what can be obtained when learning only from a narrow D_{robot}. We expect this for two reasons. First, in other domains large quantities of internet data have allowed for unexpected ‘emergent’ capabilities [212, 39]. Second, diverse internet video paired with language annotations offers a path towards stitching together the lower-level knowledge obtained from robot and video data with the rich abstractions and world knowledge that can be obtained from text data [193].

Improvements in-distribution of D_{robot}. Finally, we also expect video data to yield improvements in tasks that are in-distribution of the robot dataset. Utilising a large video dataset can allow for improved data-efficiency with respect to D_{robot} [188]. Additionally, LfV approaches may obtain higher absolute task performance (e.g., higher success rates) in settings in-distribution of D_{robot}, versus non-LfV approaches [289].

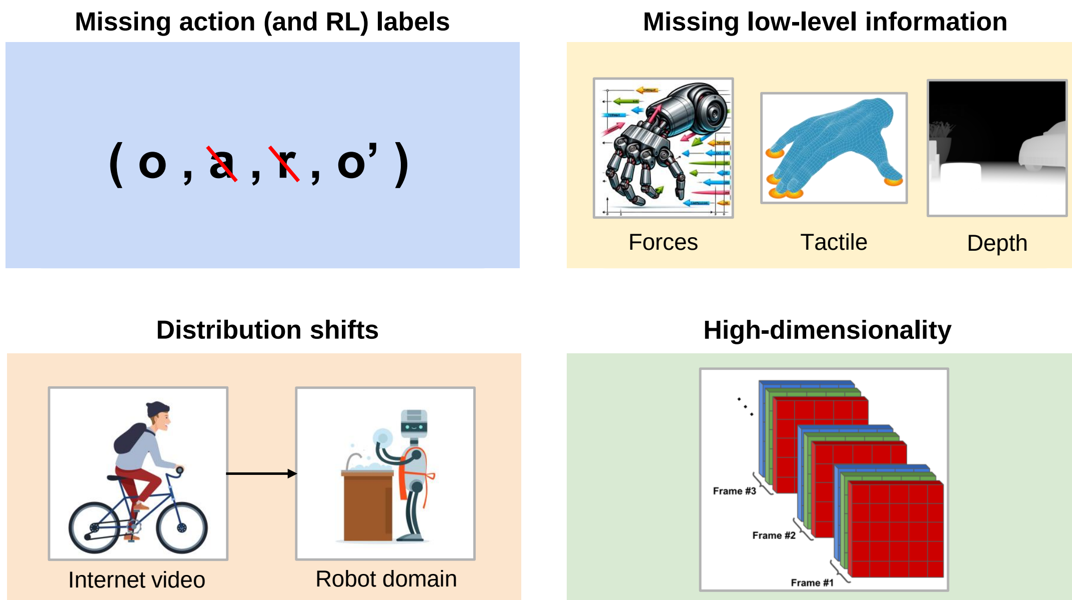


Fig. 3. **Key challenges in LfV** (see Section 3.4) are visualised, including: missing (action and low-level) information in video, LfV distribution shifts, and the high-dimensional nature of video data.

3.4 Fundamental Challenges

Learning from internet video for robotics comes with some fundamental challenges. These render LfV a non-trivial problem. Awareness of these challenges is crucial for understanding the motivation behind existing LfV research (Section 4) and the future promising research directions we highlight (Section 6). Note, none of these challenges have yet been fully solved.

Missing Action Labels. Raw video data lacks the action labels required by existing imitation [36] and offline RL [49] methods for learning from demonstrations. Moreover, adding low-level robot-action labels retrospectively to internet video data is non-trivial (see Section 4.2.1): the unrestricted action space of heterogeneous video data often cannot map cleanly to the robot action-space [310]. It is also challenging to retrospectively add other missing RL-relevant metadata to raw video, such as reward labels which can inform on the quality of the data (see Section 4.3.3), goal labels which can be useful for goal-conditioning (see Section 4.3.4), or end-of-episode labels.

Distribution-shift. There may be significant distributional shifts between an internet video dataset and the downstream robot domain. This can include differences in physical embodiments, camera viewpoints, tasks, and environments. Humans in videos may perform behaviours which are sub-optimal or irrelevant to the downstream robot. Certain useful tasks and behaviours may be underrepresented in internet video (e.g., coverage over specific factory tasks may be poor). These shifts present a challenge to deep learning methods.

Missing Low-level Information. For certain skillful or dexterous behaviours, robots require low-level percepts such as tactile sensing, forces, proprioception, or depth sensing. This crucial low-level information is not explicitly available in internet video. A key challenge in LfV is to obtain generalization beyond D_{robot} despite the missing low-level information in D_{video} (see Figure 2).

Controllability, Stochasticity, and Partial Observability. In unlabelled video, it can be difficult to disentangle which parts of a transition are affected (i.e., controlled) by a specific agent's actions or which are due to the external environment or noise [104]. This can be problematic for methods that attempt to extract action information from video (Section 4.2.1). Furthermore, the stochastic nature and partial-observability of the underlying environments in video can make accurate video prediction a challenge [13].

High-dimensionality, Noise, and Redundancy. Methods that learn from or generate video data are typically computationally demanding due to the high-dimensional nature of video data. Additionally, video can contain significant noise and redundant information. These characteristics make it challenging to extract meaningful information and representations from video data [21].

4 Methods

This section reviews methods for learning from video data for robotics. We first review video foundation models as a general-purpose means for extracting knowledge from internet video (Section 4.1), noting that advances in these approaches will directly serve advances in scalable LfV methods. We then review common methods used to tackle key LfV challenges (Section 4.2). Finally, we move to our primary taxonomy of the LfV literature, categorizing LfV methods according to which RL knowledge modality benefits from the use of video data (Section 4.3).

4.1 Extracting Foundational Knowledge from Internet Video

Internet video lacks action labels and has various distribution shifts with the robot domain. Given these LfV challenges (see Section 3.4), a key LfV question is: *How to extract robotics-relevant knowledge from large-scale, heterogenous internet videos?* The video foundation model (FM) literature offers potential answers here.

Video FMs are large deep learning models trained on large internet-scraped video datasets. From this data, they extract general knowledge and capabilities useful in a wide range of downstream settings [173]. This includes robotics-relevant knowledge related to semantics, physics, and behaviours. As such, video FMs and corresponding techniques can be highly applicable for the following LfV use-cases (see Figure 4):

- *Using pretrained models:* A pretrained video FM can be transferred or adapted (e.g., finetuned) for downstream robot applications, allowing the robot to leverage its broad world knowledge. For example, a video prediction model can be adapted into a robot dynamics model [306].
- *Using techniques and datasets:* Techniques and datasets originally used for video FMs can be customized for robotics purposes. For example, a robot FM can be trained on both video and robot data, using techniques inspired by video FMs [251]. These techniques can allow LfV to better extract representations and knowledge from internet video, thus improving the generality of the robot.

The promise of these video FM use-cases is backed up by prior work demonstrating that text and image FMs can benefit robotics. Scalable architectures transferred from these other domains have allowed robot techniques to better fit large, diverse robot data [36, 264]. Knowledge and representations from text and image FMs have been transferred to the robot domain to improve capabilities, generalization, and common-sense [6, 37, 131]. The use of video FMs may lead to further improvements due to the unique robotics-relevant information video data can offer.

Video foundation models and techniques will likely be key driver to the success of scalable LfV approaches. As such, this section provides LfV researchers with essential information regarding potential use-cases and the lower-level workings of these techniques. We loosely categorise the video foundation model literature into: video encoding methods (Section 4.1.1); video prediction methods (Section 4.1.2); and video-to-text generation methods (Section 4.1.3).

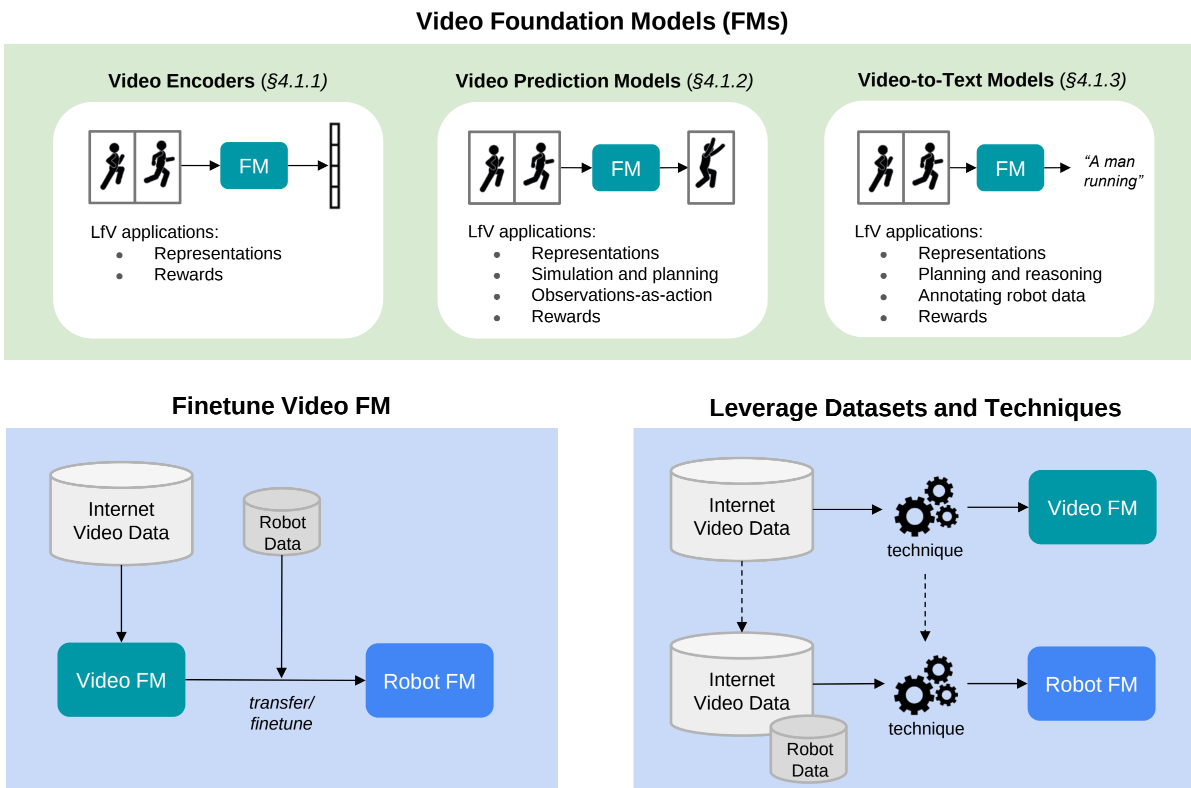


Fig. 4. **Video Foundation Modelling for LfV** (see Section 4.1). The top green box outlines different categories of video foundation models and their applications to robotics. The bottom blue boxes illustrate ways video foundation models can contribute to LfV: (left) pretrained video foundation models can be finetuned into robot foundation models; (right) video foundation model techniques and datasets can be used to train robot foundation models.

4.1.1 Video Encoders. Foundational video encoders ($p(z|\tau)$, where z is a representation and τ is a video clip) can provide rich and robust video representations for downstream robotic applications. Specifically, they can useful as follows:

- *Representations:* LfV representation transfer approaches take a pretrained video encoder and finetune it into an RL KM (e.g., into a policy, see Section 4.3.1).
- *Rewards:* LfV approaches have used frozen pretrained video representations to help define robot reward functions [77, 253, 188] (see Section 4.3.3).

Methods. We now provide a general overview of representation learning from video, before providing more details on state-of-the-art (SOTA) approaches:

- *Overview of video representation learning.* Let $\bar{X}$ be a video clip X paired with some corresponding labels and data modalities. Most methods for learning video encoders can be framed as first deriving a 'label' Y from $\bar{X}$ and then using Y to define a learning objective. Y could be some transform of the video X itself (e.g., a masked version of the video, see Wang et al. [280]). Otherwise, Y could be derived from additional

information paired with the video such as: labels (e.g., object masks and bounding boxes, see Ding et al. [64]); meta-data [223]; or corresponding data of another modality (e.g., language annotations, see Grauman et al. [87]). Once Y is obtained, there are a number of ways it can be used. Prediction objectives predict Y from X. For example, predicting the next frame from the previous frame (i.e., video prediction) [78, 116]. Joint-embedding approaches first embed X and Y before calculating the loss [192, 297, 21]. Here, contrastive learning approaches have been heavily explored [223].
- *SOTA Approaches.* The literature has proposed a number of approaches for learning rich spatio-temporal representations from video in a *scalable* manner. Video-text losses leverage text annotations to learn semantic representations from video. This can include: video-text contrastive losses [297, 327, 196, 153, 284], video-text matching [153], masked language modelling [153], and video-to-text losses [301, 196]. Auto-encoding approaches have also proven popular. State-of-the-art methods have used masked auto-encoding (MAE) [269, 280, 83] and vector-quantized auto-encoding (VQ-AE) [233, 302, 313, 275, 40]. A number of works have explored the use of student-teacher distillation losses [281, 327, 153]. Joint-embedding prediction architectures (JEPA) have been proposed to mitigate issues related to video's high-dimensionality and noise in video [21].

4.1.2 Video Prediction Models. This section is primarily concerned with models that can perform next-frame video prediction: $p(o_{t+1}|o_{t-k:t})$. Also relevant are models that can more generally perform some form of conditional video generation: $p(\tau|c)$ (where τ is a video clip and c is some conditioning information) [308]. Via training on internet video, such models can learn information regarding the world dynamics and human behaviours. They can thus be useful for downstream robotic applications in the following ways:

- *Dynamics:* A video prediction model can be adapted into a robot dynamics model $p(o_{t+1}|o_{t-k:t}, a_t)$ to serve as a planner [68] or simulator [306] (see Section 4.3.2).
- *Policies:* The video prediction objective implicitly models the distribution of behaviours in the video dataset [74]. As such, video prediction models can act as policies by generating 'observations-as-actions', videos of future behaviours the robot should execute [67] (see Section 4.3.1).
- *Representations:* Due to the relevant information they represent, video predictors can be used for LfV representation transfer approaches [289].
- *Rewards:* Finally, a reward signal can be defined that encourages the robot to match the behaviour expected by the video predictor [74] (see Section 4.3.3).

Methods. Diffusion models, autoregressive transformers, and masking transformers have proven to be particularly effective and scalable architectures in recent years.

- *Diffusion models* can easily model continuous output spaces and can sample multiple frames in parallel. However, their sampling speeds can be slow and long video generation remains a challenge [308]. Computational efficiency can be aided by performing the diffusion process in a learned latent space [38, 20, 33] rather than in pixel space [100, 247, 101]. Many diffusion video prediction models augment their training data by making use of image data (which can come in higher quality and quantity than video data) [100, 101, 20, 81, 32, 56]. The diffusion video prediction pipeline can often involve several intermediate steps, including: key-frame generation, interpolation between key-frames, and spatial super-resolution upsampling [101, 81, 332]. However, Bar-Tal et al. [20] generate the entire temporal duration in a single forward pass, obtaining improved global temporal consistency.
- *Autoregressive and masking transformers* tend to make predictions in a learned latent-space [302, 312, 80, 136]. Predicting in latent space can improve computational efficiency and can mitigate issues related to pixel-level noise and redundancy [302]. The latent space is generally pretrained via vector-quantized auto-encoding (VQ-AE), providing a discrete token-space suitable for the transformer architecture. Versus

autoregressive methods [302, 105, 40], masking transformers train the model to decode masked tokens in parallel [313, 317, 91]. This is usually achieved via the use of MaskGIT decoding [44]. Masked models are more computationally efficient and do not suffer from the 'drifting' effect that can occur in auto-regressive models [308].
- *Conditioning information* can simplify the video prediction problem and allow for more control over generated videos. Such conditioning is valuable for downstream robotics: it can allow us to simulate the effects of different 'action' strategies. Yang et al. [308] outline various popular conditioning schemes for video generation models $p(\tau|c)$. In particular, language conditioning can allow for flexible and intuitive control over generated video, at varying levels of detail [38, 20, 136]. Amongst other possibilities, video generation can also be conditioned on future images (i.e, video in-filling, see Höppe et al. [102]) or on action representations [40]. We outline various action representations that may be suitable here in Section 4.2.1.

4.1.3 Video-to-Text Models. This section details to models with video-to-text capabilities, $p(\text{text}|\tau)$. A capable video-to-text model can perform, for example, video question-answering or video summarization. Such a video-to-text foundation model could be valuable to robotics in the following ways.

- *High-quality representations:* A capable video-to-text model will have robust, high-quality video representations. Compared to a video-only model, it will have improved high-level semantic representations. Compared to an image-to-text model, it will have improved temporal-dynamic representations. Robotics can bootstrap from such models via representation transfer [37], or by adding robot data directly into the model's pretraining corpus [217].
- *Grounded reasoning and planning:* LLMs have proven useful as planning modules in robotics [6], but their lack of grounding in the physical world is limiting. In contrast, video-to-text models can perceive the environment through information-rich video, allowing for improved and closed-loop reasoning and planning.
- *Annotating robot data:* High-quality language annotations can provide valuable conditioning information in many ML domains [24, 38]. Robotics is no different [264]. Capable video-to-text models could serve as useful language annotators for robotic datasets [31].
- *Rewards:* A sufficiently capable video-to-text model can provide reward or value estimates for a robot learner. For example, this could be achieved through a visual-question answering framework [69], or via an RL-from-AI-feedback framework [133] (see Section 4.3.3).

Methods. We now give an overview of existing video-to-text methods and models. We note that research here is somewhat preliminary: low-quality video captions and the difficulties of the video modality means progress lags behind foundation models in other domains.

- *Compositional approaches* have often been used to obtain video-to-text pipelines [51, 240, 323, 155]. For example, Chen et al. [51] use an image-language model to answer questions about individual video frames, and an LLM to synthesize this information to produce a global summary. However, the lack of end-to-end video training in compositional approaches mean they can lack rich, nuanced video representations.
- *Leveraging pretrained LLMs via adaptors and finetuning.* Approaches here typically first obtain a pretrained LLM and an (often) pretrained video encoder, and then define an adaptor module to channel information from the video encoder output into the LLM [160, 172, 325, 196]. Once the new video-to-text architecture is defined, a common training scheme is: (i) convert large, diverse video-text data into token sequences and perform next-token-prediction pretraining, then (ii) perform supervised instruction-tuning on small, high-quality instruction datasets [160, 325]. Following trends in LLMs, future work could investigate a third RLHF finetuning stage [128]. During training, the LLM and video encoder may be finetuned [160], or kept frozen [172].

- *Natively multi-modal models.* Previously discussed methods have involved combining and finetuning models not initially intended for video-to-text purposes. Recent multi-modal any-to-any sequence models—where video-to-text is formulated as an interleaved video and text sequence modelling problem—represent a step towards more *natively* multi-modal models. Liu et al. [163], Team et al. [263], and Jin et al. [119] all train multi-modal any-to-any (or any-to-text) autoregressive transformers via next token prediction.

4.1.4 Challenges. Scaling datasets and model sizes has recently led to impressive advances in video ML [327, 306, 285]. Most notably, the recent Sora model [306] demonstrated significant improvements in video prediction visual quality, physical realism, and generation lengths. However, there are still major limitations in the capabilities of current video foundation models. Models can hallucinate, and can struggle with spatial relationships, fine-grained spatio-temporal details, and long-term understandings and generation [160, 172, 306].

One bottleneck to further progress here is the quality (and quantity) of available video-data (see Section 5.1). In particular, improved annotations with precise low-level details will enhance robotics-relevant fine-grained understandings. The computational demands of processing high-dimensional and long video data present is another challenge. Improved model architectures can help here [308]. For example, progress in long video understanding will be aided by efficient architectures for handling longer contexts [164, 88, 19]. Other directions to improve video foundation models include using RL finetuning to address hallucination issues [29], and leveraging 3D information to improve the physical realism of video generations [331]. Improved methods for fine-grained, single-step conditioning of video predictions will be relevant to robotics [40]. Finally, we advocate for progress in open-sourced foundational video prediction models: open-sourced models make LfV research more accessible to the wider community.

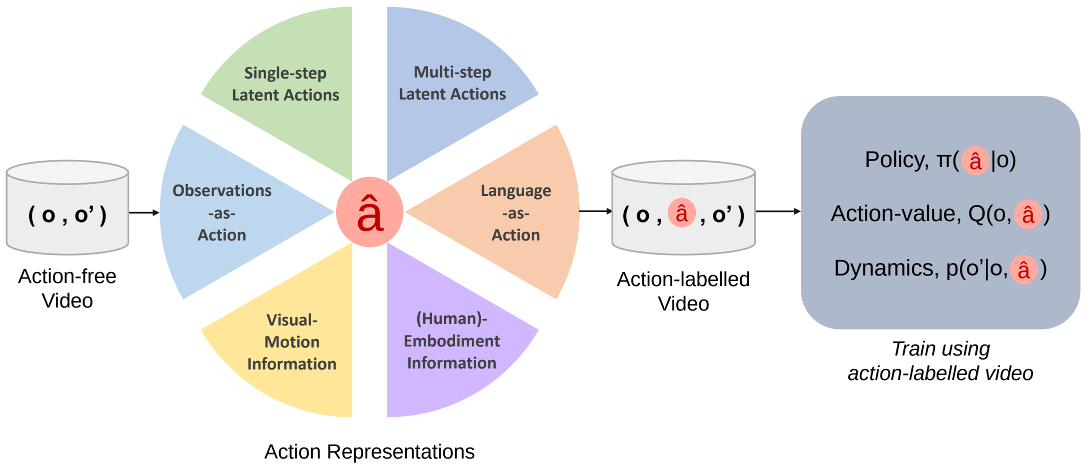
4.2 Video-to-Robot Transfer: Addressing Missing Action Labels and Distribution Shift

The fundamental challenges associated with LfV are outlined in Section 3.4. In this section, we detail two categories of techniques that each address a fundamental challenge: (1) The use of action representations to mitigate missing action labels in video data (Section 4.2.1). (2) The use of representations designed to explicitly address LfV distribution-shift issues (Section 4.2.2). These techniques are often used as a single component within a larger LfV pipeline. We thus describe these technique in isolation here, in advance of our main analysis of the LfV literature in Section 4.3.

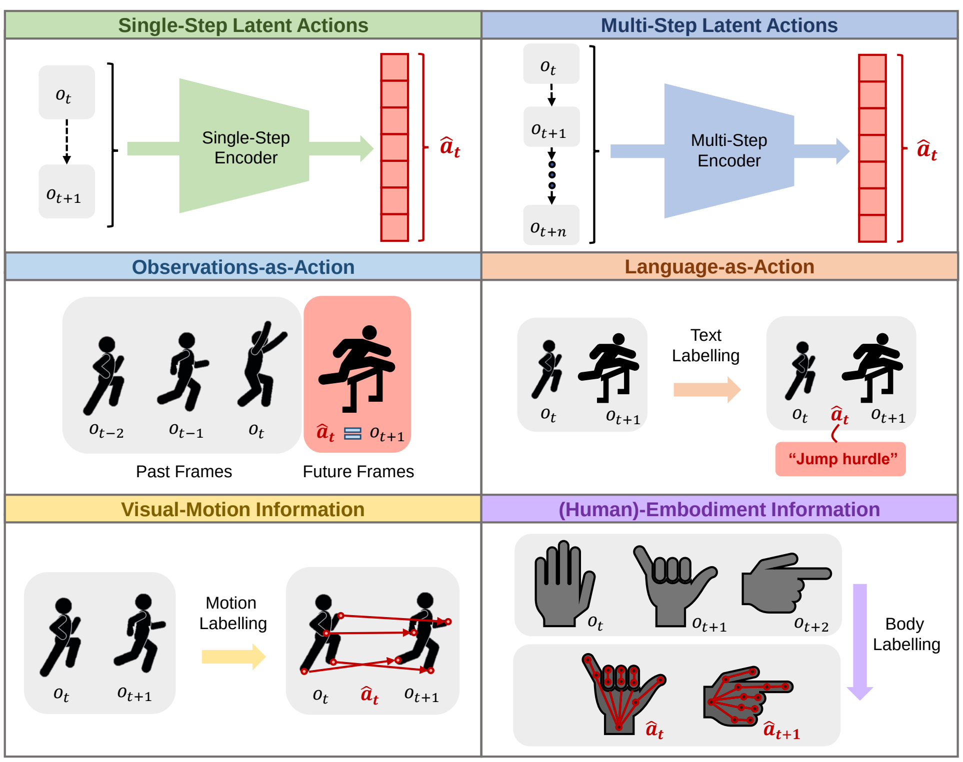
4.2.1 Action Representations. Video data transition tuples are missing action labels: they come in the form of (o_t, o_{t+1}). In this section, we review works that define or learn some representation space that is analogous to the notion of an action; an *alternative action representation.* We use $\hat{a} \in \hat{\mathcal{A}}$ to denote a single alternative-action and its underlying alternative-action space. With $\hat{\mathcal{A}}$, video data can be relabelled from (o_t, o_{t+1}) to $(o_t, \hat{a}_t, o_{t+1})$ and used to train an 'alternative-action' version of an RL KM (see Figure 5a). Alternative-action policies $\pi_{\text{alt}}(\hat{a}_t|o_t)$ [227], dynamics models $p_{\text{alt}}(o_{t+1}|\hat{a}_t, o_t)$ [40], and value functions $Q_{\text{alt}}(o_t, \hat{a}_t)$ [27] have all been previously trained. These alternative-action KMs can be useful downstream if a mapping to the robot action-space, $f : \mathcal{A} \to \hat{\mathcal{A}}$, is obtained [278, 288, 67]. Otherwise, their representations can prove highly useful [27, 227].

There are a number of characteristics we desire from our action representations $\hat{\mathcal{A}}$:

- $\hat{\mathcal{A}}$ *should be inferable from video data.* Action representations that can be obtained from raw, unlabelled video are appealing as they can leverage the full range of available video data. However, $\hat{\mathcal{A}}$'s that require labelled video (e.g., language captions) are also valid.
- $\hat{\mathcal{A}}$ *should contain 'action' information.* We can operationalize this in two ways. First, $\hat{\mathcal{A}}$ should be predictive of future video frames: learning $p_{\text{alt}}(o_{t+1}|\hat{a}_t, o_t)$ should be easier than $p(o_{t+1}|o_t)$. Second, $\hat{\mathcal{A}}$ should be predictive of robot actions: learning $\pi(a_t|\hat{a}_t, o_t)$ should be easier than $\pi(a_t|o_t)$.



(a) Action-free videos can be labelled with alternative-action representations $\hat{a}$. This labelled data can be used to train an alternative-action RL knowledge modality (e.g., a policy, value function, or dynamics model).



(b) Categories of action representations that can be learned or obtained from video data.

Fig. 5. **Recovering action representations from video** to overcome the missing action label problem in LfV (Section 4.2.1).

- $\hat{\mathcal{A}}$ *should be transferable across embodiments.* It should capture a high-level, general notion of an action, and transfer from the embodiments in the video data (i.e., humans) to the target robot embodiment(s).
- $\hat{\mathcal{A}}$ *should be well-structured.* It may be beneficial for $\hat{\mathcal{A}}$ to contain minimal information, to be well disentangled, and to maintain consistent meaning across the state space. This can allow $\pi(\hat{a}_t|o_t)$ and $\pi(a_t|\hat{a}_t, o_t)$ to be learned in a more data-efficient and generalizable manner [221, 40].

We now outline the main categories of $\hat{\mathcal{A}}$ we have identified in the literature (see Figure 5b). We focus on detailing how $\hat{\mathcal{A}}$ can be defined and learned from video data. More details on how $\hat{\mathcal{A}}$ can be used downstream for robotics are found throughout Section 4.3.

Single-step Latent Actions. Here, we consider *learned* representations where $\hat{a}_t$ (i.e., the latent action) only contains information regarding the action that took us from o_t to o_{t+1}. Such a latent representation is commonly learned using a next-observation prediction objective $p(o_{t+1}|\hat{a}_t, o_t)$, such that $\hat{a}_t$ is informative for the predictions of a forward dynamics model (FDM) [72, 221, 227, 40, 310]. A trainable (latent) inverse dynamics model (IDM) can be used to encode past-frames and infer $\hat{a}_t$ [221, 40]. Here, some form of regularisation, such as a vector-quantization bottleneck [227, 40], can be employed to prevent the IDM from copying the entire observation o_{t+1} into the latent action.

We note some potential limitations of these approaches. First, these latent actions model all visual changes that occur in a video. Ideally, $\hat{a}_t$ would only represent changes due to actions taken by a single agent. Second, as these approaches model environment transitions on a visual level, thus they may omit important non-visual low-level information (such as forces). We finally note that more research into learning single-step latent action-spaces from realistic internet video is required.

Multi-step Latent Actions. This refers to *learned* action representations, $\hat{a}$, that contain information regarding the actions taken across *multiple* time-steps of a video (i.e., from o_t to o_{t+k}). Any means of representing video segments could be applicable here, including the video representation methods seen in Section 4.1.1. Here, we only detail methods used directly in the LfV literature. Building upon the 'latent plans' literature [168, 55, 220], Wang et al. [278] learn latent plan representations from action-free videos via variational auto-encoding. However, they use 3D human hand trajectories as the decoder target (rather than raw video). Fan et al. [77], Lifshitz et al. [159], and Sontakke et al. [253] use video-language contrastive losses on language-captioned video. Chane-Sane et al. [43] and Chen et al. [50] perform supervised contrastive learning on video labelled with the action being performed. Xu et al. [298] use a self-supervised learning clustering approach to learn video representations. Further approaches for learning multi-step latent actions are seen in the works of Tomar et al. [268], Pertsch et al. [204], and Cai et al. [42].

Observations-as-action. Future observations (or an encoding thereof) provide information regarding what actions will be taken next in a video. Thus, they can be used as action representations. Observations-as-actions can be implemented on varying time-horizons, as we now describe. (1) *Next-observation-as-action:* These methods use the next observation o_{t+1} as the $\hat{a}_t$ label [67, 265, 95]. o_{t+1} provides clear information regarding the action that should be taken at o_t. (2) *Observations-as-subgoals:* A subgoal can be thought of as a high-level action. In 'observations-as-subgoals' methods [30, 199, 27], an observation (or embedding thereof) from k time-steps into the future is used as a sub-goal / action representation. Some simple strategies for defining sub-goals in the video data include: choosing a fixed time-horizon k [30, 67], or randomly sampling observations beyond the current timestep [27]. More complex strategies include using key-frame identification to identify bottleneck states in video [205, 165].

Language-as-action. Natural language can be used as a flexible, high-level action-space (e.g., $\hat{a}_t$ = "pick up the cube") [23, 244, 293], and can allow for interfacing with other language models [68]. Some video datasets come with language-action annotations [87]. Otherwise, standard annotations can be further processed (e.g., using LLMs) [185]. If the video does not come with any language annotations, manual or automated captioning methods can be used (see Section 5.1.2). One downside here is that language is coarse and may omit important lower-level action information.

Visual Motion Information. Other works have used visual motion information in video to define $\hat{\mathcal{A}}$. Wen et al. [288] use 2D point trajectories, obtained by tracking points on objects throughout the video via an off-the shelf point tracker [124]. Yuan et al. [318] similarly use 3D point trajectories obtained from 3D annotated datasets. Elsewhere, Ko et al. [134] use an off-the-shelf model [296] to predict optical flow, giving a pixel-level dense correspondence map between two frames. Nasiriany et al. [191] represent actions via visual arrows in images. Finally, Wang et al. [279] use structure-from-motion [228] to recover action information, Yuan et al. [319] use the motions of object-centric representations, and Yang et al. [306] optionally use camera frame motion information as conditioning information for a video predictor.

(Human)-embodiment Information. Off-the-shelf human-hand detection models [219, 239] can extract hand poses or affordances from videos which can act as an action representation [26, 15, 243, 210, 211]. For example, $\hat{a}_t$ can be defined as the pose that should be reached at time $t + 1$. Bharadhwaj et al. [26] use object masks in addition to human poses to define $\hat{\mathcal{A}}$. Animal embodiment information could also be used [202]. More details on methods for detecting human hand poses and affordances from video are provided in Section 4.2.2.

IDM Pseudo-action Labels. Though not technically an *alternative* action representation, these methods train an inverse dynamics model $p^{-1}(a_t|o_t, o_{t+1})$ on action-labelled robot data, and use it to provide pseudo-action labels for action-free video data [18, 270, 225]. However, these approaches are unlikely to scale to diverse internet video as they require either: (i) minimal domain-shift between the video data and the robot domain (e.g., Baker et al. [18] assume identical embodiments), or (ii) an explicit mechanism to deal with domain-shift that may not scale well to diverse internet video [225, 130] (see Section 4.2.2).

Discussion. We now discuss key considerations regarding the feasibility and utility of alternative action representations for LfV, and possible directions for future work:

- *Can $\hat{\mathcal{A}}$ be easily obtained from internet video?* Observations-as-action can be obtained from any raw video and can (in theory) leverage the full range of available video data. Single-step latent actions, and (some) multi-step latent actions can also be obtained from raw video, but an additional learning step (that may face optimisation difficulties) is required. Language-as-action, visual motion information, and (human)-embodiment information all require labelled video. Nevertheless, often off-the-shelf models can provide labels, and curated video datasets often come paired with language annotations (see Section 5.1.3). Visual motion information and (human)-embodiment information approaches may require higher degrees of structure in the video (e.g., a relatively fixed frame); it is unclear how sensible these representations will be when applied to heterogenous internet video data.
- *Once obtained, how useful is $\hat{\mathcal{A}}$?* The information content of a given $\hat{\mathcal{A}}$ will determine how useful it will be. A longer-horizon or more high-level $\hat{\mathcal{A}}$ (e.g., language-as-actions) may ultimately aid downstream performances in long-horizon tasks. However, shorter-horizon $\hat{\mathcal{A}}$'s with more low-level information may be more informative for decoding a_t from $\hat{a}_t$. Note, video-based actions representations will not usually contain all information required to accurately decode the exact low-level robot action. As such, in practice, any decoder may need to be defined as $\pi(a_t|\hat{a}_t, o_t)$ rather than $\pi(a_t|\hat{a}_t)$.
- *Future directions.* There is no work comprehensively comparing the utility of different action representations for LfV. We advocate for empirical research to help answer the questions we have posed above in the two paragraphs above.

4.2.2 Representations to Address LfV Distribution-Shift. Distribution shift between internet video and the target robot domain poses a challenge to LfV approaches (see Section 3.4). Such distribution shifts can hinder our ability to transfer knowledge from the video data to the robot. A line of LfV research has attempted to overcome these issues by explicitly designing representations of (human) video that are more transferable to the robot domain. Once the transferable representation is obtained, works have most commonly used it to: (i) define a reward function encouraging the robot to match the behaviour in the video [321], or (ii) to help train a policy via behaviour cloning of the represented video [26].

In this section, we focus solely on describing methods for obtaining representations that explicitly address LfV distribution-shift. Details regarding how the obtained representation can be used downstream to aid robot learning are found throughout Section 4.3.

Method: (Human)-embodiment-aware Approaches. In LfV, we are most often interested in learning from human behaviour; in particular in replicating the effects of the human hand. There is a line of LfV research that explicitly detects human-embodiment information in video, and subsequently transfers this information to the robot. Note, animal embodiment information can also be used [202].

- *What types of human-embodiment information can be detected?* (1) Poses: Several works explicitly estimate the pose of the human body or hand in videos [243, 248, 151]. This involves estimating the positions and orientations of various joints or key points on the body via off-the-shelf models. Once a human-hand pose is detected, it may need to be retargeted to the robot embodiment. This can be achieved by directly optimising a loss function at inference [243], or by training a retargeting network to minimise a retargeting loss function [248]. (2) Affordances: A common affordance seen in the LfV literature is the combined use of grasp points and grasp trajectories [15, 180] – this abstraction transfers naturally from human to robot embodiment. This affordance information can be extracted from videos using an off-the-shelf model [239], and using further tricks to ensure accurate, usable affordance labels [15]. Note, 2D affordances can be converted to 3D using depth estimation [180]. (3) Masks and Bounding boxes: Masks or bounding boxes of human hands and objects can be used [26]. These can be obtained from videos using off-the shelf models [132, 326, 151] or via labelled datasets [60].
- *How to use this human embodiment-centric information?* A video dataset labelled with the human-centric information described above can allow for one of the following to be performed: (1) The original process for extracting the human-centric information from the video/image can be distilled into a single model [180]. This can be useful when the original method is convoluted or unlikely to generalize. This process may also be a good auxiliary representation learning objective for robot manipulation [15]. (2) The extracted embodiment information can be treated as an 'alternative action representation' [26, 15, 243, 210, 211] (see Section 4.2.1), or as a state-observation when retargeted to the robot embodiment [214].
- *Limitations.* There can still be difficulties transferring this human-embodiment information to the robot. For example, a model trained to propose future poses solely on human images is unlikely to generalise zero-shot to robot images. This has led to tricks being used in the LfV literature, such as predicting affordances only when the embodiment is out of frame [15], or in-painting human embodiments over robot embodiments [26].

Method: Learned Invariant Representations. These are methods that *learn* a representation that is invariant to a specific distribution shift between the video dataset and robot domain.

- *Approaches.* (1) Domain confusion techniques can learn representations that are invariant across viewpoints [254] and embodiments [225]. (2) Contrastive learning and related techniques can encourage invariance across a particular axis; such as viewpoints [234] and visual changes [12]. (3) Temporal cycle-consistency objectives can learn embodiment invariant representations Zakka et al. [321]. (4) Factorized representations learn one representation that is common across the video and robot distributions, and one that is unique to each distribution [226, 241]. Similarly, Chang et al. [47] use embodiment segmentation and in-painting removal to obtain explicit factorized representations of the agent and environment. (5) Image translation methods can convert an image from the source distribution (e.g., human embodiment) to the target (e.g., robot embodiment) [250, 14, 26, 154].

- *Limitations.* (1) Invariant representations can omit useful information, such as important differences between human and robot embodiments. (2) These methods can have overly strict requirements on the video data. For example, Sermanet et al. [234] assume access to multiple videos of the same scene but from different viewpoints. (3) Many of these methods only provide invariance to a single type of distribution shift (e.g., embodiment differences only). In reality, there may be many distinct shifts between any given video and the robot's setting.

Method: Transferable Abstractions. Some methods exploit abstractions that naturally transfer well from human to robot. This includes: object-centric graphical representations [245, 140]; object-centric activity-context priors [186]; key-points in video [126, 295]; and embodiment-agnostic affordances [15, 180]. These methods all respectively benefit from the use of off-the-shelf object, key-point, and human-hand detectors. Language has also been used as a transferable abstraction [50, 185, 204].

Discussion. These methods highlight the trade-off that must be made between: (i) imposing structure and inductive biases to improve performance in narrow settings, and (ii) opting for end-to-end learning methods that are more scalable, but are less effective in the small data regime and narrow settings. The central focus of this survey is on scaling to diverse internet data in order to tackle unstructured environments with generalist robots. As such, we advocate for methods following the spirit of (ii). Through this lens, we regard some of the methods in this section to be less promising, and have commented on their limitations above.

4.3 Robot Learning and Reinforcement Learning from Video

This section presents our main analysis of the LfV literature. We taxonomize this literature according to which downstream RL knowledge modality (KM) most directly benefits from the use of video data by the LfV method (see Figure 6). The KMs we consider here are: policies (Section 4.3.1), dynamics models (Section 4.3.2), reward functions (Section 4.3.3), and value functions (Section 4.3.4). For brief descriptions of each of these KMs, we refer the reader back to Section 3.1.

4.3.1 Policies. Ultimately, the goal of LfV is to use D_{video} to help obtain a policy $\pi(a_t|o_t)$. In this section, we detail literature where π is the RL KM that most directly benefits from the use of video data. We first identify and detail several distinct categories of methods (see Figure 7), before moving to a brief discussion.

Method: Representation Transfer. These methods pretrain visual representations on video data using some learning objective. Subsequently, the representations are transferred (either frozen or to be finetuned) to aid downstream learning of a policy. Figure 7 visualizes this process. There have been several large-scale analyses of representation transfer from video data [120, 246, 328, 41, 108, 174].

A number of learning objectives have been explored in the literature. This includes: frame-level objectives such as masked-autoencoding (MAE) [213], contrastive learning [52, 120], and others [328]; time-contrastive learning objectives [171, 170, 188]; video prediction objectives [289, 95, 116]; temporal-difference learning [27]; predicting affordances from videos [15]; predicting latent actions [227]; and objectives that leverage language labels, such as image-language contrastive learning [170], video-language alignment [188] and video captioning [125]. Works have often combined multiple objectives [188, 170, 125].

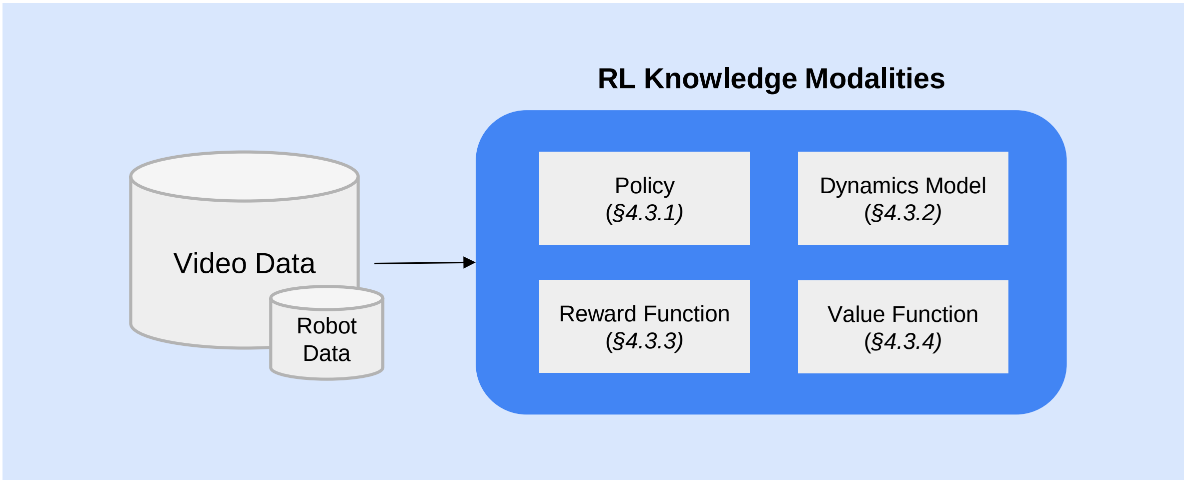


Fig. 6. **Leveraging video to benefit RL and robot learning** (Section 4.3). We categorise the LfV literature according to which Reinforcement Learning (RL) Knowledge Modality (KM) benefits from the use of video data, resulting in the taxonomy illustrated in the above figure.

The use of video data relevant to the robot is important [120]. Nair et al. [188], Ma et al. [170], Karamcheti et al. [125], and Majumdar et al. [174] leverage ego-centric human and/or navigation video. Radosavovic et al. [213] leverage third-person video of humans interacting with objects [239, 181]. Others have leveraged videos of robots [61], or first-person videos of humans operating robot-like grippers [237]. Combining datasets to improve data diversity has proven beneficial [174, 61, 213].

Method: Multi-modal Models. These methods train a monolithic model jointly on video and robot data. Reed et al. [217] train a sequence model on multi-modal data, including image, text, and robot data, but do not use video data. Radosavovic et al. [214] train a sequence model for locomotion on both action-labelled robot data and action-free pose trajectories extracted from YouTube videos. Sohn et al. [251] present an 8 billion parameter any-to-any transformer trained on text, images, videos, and robot perceptions and actions (but experimental details are not published). These recent methods may obtain positive transfer by learning from different modalities of data, and are promising due to their simplicity and scalability.

Method: Alternative-action Policies $\pi_{alt}(\hat{a}_t|o_t)$. In Section 4.2.1, we outlined different alternative-action representations $\hat{\mathcal{A}}$. Here, we detail methods that use such representations to train a policy $\pi_{\text{alt}}(\hat{a}_t|o_t)$ from video data. Such a pretrained $\pi_{\text{alt}}(\hat{a}_t|o_t)$ can be useful for representation transfer, or to condition a decoder $\pi(a_t|\hat{a}_t, o_t)$ (see the following paragraph [227]).

Works have trained a $\pi_{\text{alt}}(\hat{a}_t|o_t)$ using: single-step latent actions [72, 227, 40]; multi-step latent actions [278, 204]; language-actions [7, 306, 68, 185]; observations-as-action, including next-observation-as-action policies [67, 265, 95] and sub-goal policies [30, 199]; visual motion information [288, 318]; and human-embodiment information [15, 243, 210, 211, 203, 26].

A common training process involves assuming the video data contains relatively expert behaviour, labelling the video data with the alternative-action, and training $\pi_{\text{alt}}(\hat{a}|o)$ using a supervised behaviour cloning objective [227, 40, 278, 68, 30, 288, 26]. When the behaviour in the data is suboptimal, language [68, 288] or goal-conditioning

[278] can be beneficial. Offline RL is another option when the data is suboptimal (though this is currently underexplored).

To obtain language-as-action policies, Yang et al. [306], Du et al. [68], and Mu et al. [185] supervised finetune internet-pretrained VLMs or LLMs. Black et al. [30] supervised finetune an image-editing diffusion model to edit the current observation into a subgoal image.

Method: Alternative-action Decoders $\pi(a_t|\hat{a}_t, o_t)$. Several works train a low-level robot policy $\pi(a_t|\hat{a}_t, o_t)$ that decodes $\hat{a}_t$ to a_t. This decoder can be useful when $\hat{a}_t$ is: (1) obtained from a $\pi_{\text{alt}}(\hat{a}_t|o_t)$ – i.e., *hierarchical conditioning via $\pi_{alt}(\hat{a}_t|o_t)$* [227, 67, 288]; or (2) $\hat{a}_t$ is a video instruction, usually a representation of a human demonstration video – i.e., *video-as-instructions* [43, 42, 159, 298]. The decoding policy $\pi(a_t|\hat{a}_t, o_t)$ can either be learned with data, or can be manually crafted.

- *Learning the decoder*: The D_{robot} can be labelled with alternative actions to give tuples of the form $(o_t, a_t, \hat{a}_t, o_{t+1})$, and the decoder can be trained via supervised learning on these tuples. This has been done with single-step latent actions [227, 40], multi-step latent actions [278], point-trajectory latent actions [288], next-observation-as-actions [67] and sub-goals-as-actions [30]. $\pi(a_t|\hat{a}_t, o_t)$ can also be trained via online RL with the compositional policy, $\pi(a_t|\pi_{\text{alt}}(\hat{a}_t|o_t), o_t)$, being trained end-to-end (as opposed to keeping π_{alt} frozen) [227]. Jain et al. [114] end-to-end learn a mapping from human demonstrations to robot actions, but require a dataset of human video demonstrations paired with equivalent robot trajectories.
- *Manually defining the decoder*: Ko et al. [134] infer low-level robot actions from optical flow, Yuan et al. [318] infer actions from 3D flow predictions, while Nasiriany et al. [191] map from robot action to visual arrows and back. Other works have retargeted human-hand pose actions to robot poses [211, 248].

Method: Policy-as-video. These methods are a distinct sub-section of the alternative-action methods mentioned above. They use observations-as-actions in a 'hierarchical conditioning via $\pi_{\text{alt}}(\hat{a}_t|o_t)$' scheme. A common approach here involves using a language-conditioned video predictor to generate plausible video trajectories that complete a language-specified task, before decoding actions from the generated video via an action-decoding inverse-dynamics model (IDM) $p^{-1}(a_t|o_t, o_{t+1})$ trained on D_{robot} [67]. To improve upon this scheme: Ajay et al. [7] enforce compositional consistency between LLM plans, video generations, and the IDM actions; and Du et al. [68] improve action decoding model via the use of goal-conditioned behaviour-cloning.

Method: IDM Pseudo-actions. These methods train an IDM $p^{-1}(a_t|o_t, o_{t+1})$ on action-labelled robot data and use it to label video data with pseudo-actions. The pseudo-labelled video data can then be used to help train a policy $\pi(a|o)$ [18]. However, as mentioned in Section 4.2.1, naive IDM pseudo-action approaches are unlikely to scale to diverse internet videos.

Discussion. Learning policies from videos has shown potential. Though it comes with some cons as well as pros, there are several promising directions for future research.

- *Performance gains:* The methods above have contributed to improvements in task-success rates [213, 188, 197], robot data efficiency [188, 227, 311], and generalization beyond D_{robot} [67, 278, 30, 265, 328, 161]. Nevertheless, this has often been in toy (non-generalist) settings or with toy datasets [227, 288, 265]. Gains when using diverse human video datasets have often been more modest [67, 243, 30].
- *Pros and cons:* A robot policy is ultimately what we want, and initial results suggest this is a highly promising KM to target. However, there are some disadvantages to note. Policies must output low-level actions and so the missing low-level information in video will present a challenge. Additionally, policies prefer expert data, but the behaviours in internet video may often be diverse and non-expert.
- *Future directions:* Jointly pretraining multi-modal foundation models on both video and robot data is still underexplored. Representation transfer approaches may be further improved by: training encoders that fuse

spatio-temporal information (rather than embedding individual images separately); using representations from video foundation models (see Section 4.1); or training on larger internet datasets (see Section 5.1.3). Methods that leverage representations $\hat{\mathcal{A}}$ will be limited by the scalability of their $\hat{\mathcal{A}}$ (as discussed in Section 4.2.1). An open question is whether $\hat{\mathcal{A}}$'s are best utilised explicitly in a hierarchical setup, or if

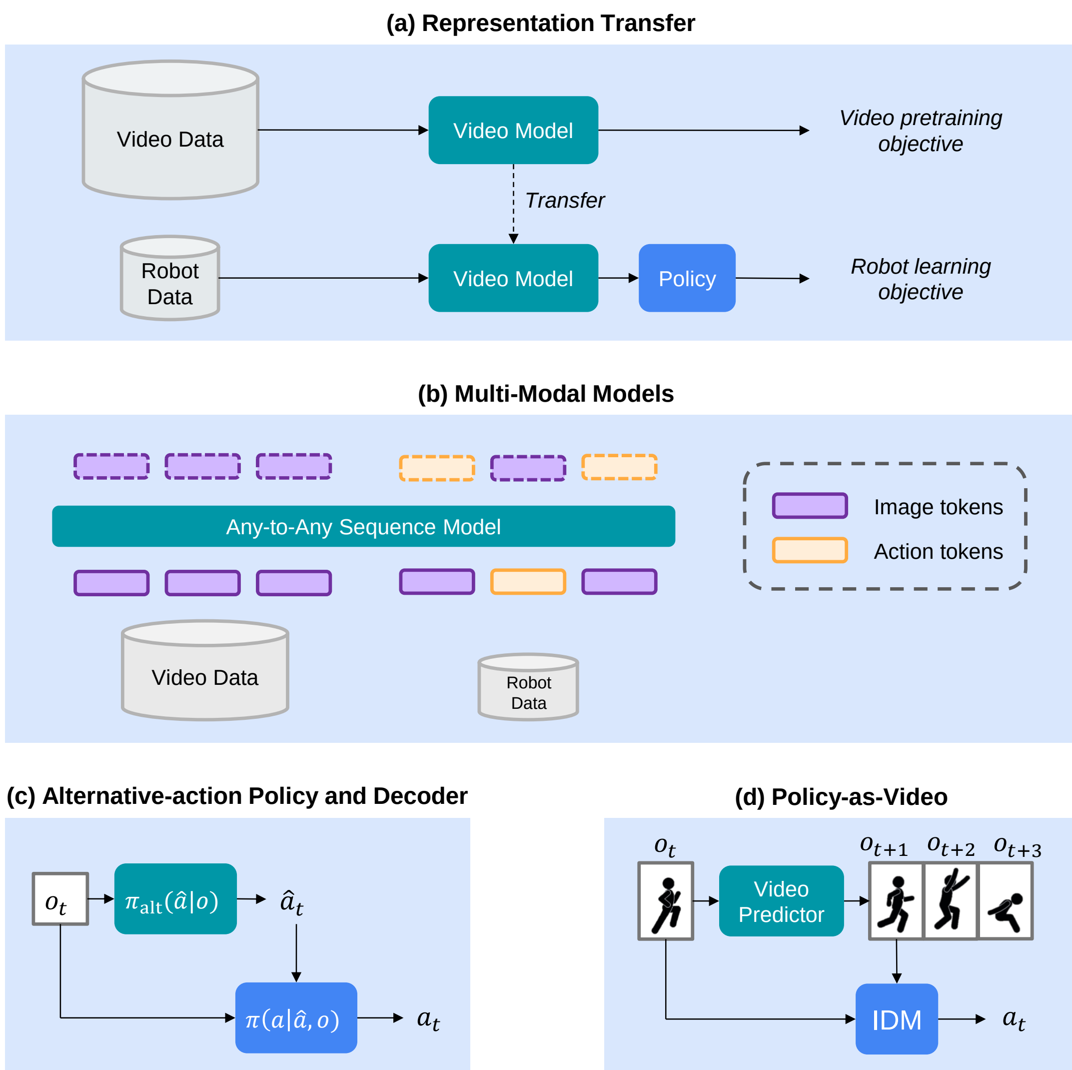


Fig. 7. **Learning policies from video** (Section 4.3.1). (a) Representations can be pretrained from video data and used downstream whilst training a policy on robot data. (b) Models can be trained jointly on video and robot data to predict robot actions. (c) A π_{alt} can be trained on video data to output an action representation (see Section 4.2.1) which conditions an action-decoding π trained on robot data. (d) Policy-as-video is an specific instantiation of this hierarchical setup where the action representation is a video and the decoder is (often) an inverse dynamics model (IDM).

they are most useful for representation transfer. In hierarchical setups, where the policy is defined as $\pi(a_t|\pi_{\text{alt}}(\hat{a}_t|o_t), o_t)$, we should explore the extent to which either π_{alt} or π is the performance bottleneck.

4.3.2 Dynamics Models. Video prediction models, $p(o_{t+1}|o_t)$, capture temporal dynamics information – information about the dynamics and physics of the world. A line of LfV research has sought to use video prediction objectives on a D_{video} to aid the learning of a robot dynamics model $p(o_{t+1}|o_t, a_t)$. This can involve pretraining $p(o_{t+1}|o_t)$ on D_{video} and adapting it into $p(o_{t+1}|o_t, a_t)$ using D_{robot}. Alternatively, it could involve jointly pretraining on D_{video} and D_{robot}. In this section, we detail the various video prediction pretraining schemes seen in the LfV literature, before outlining how the pretrained models can be adapted and utilized for downstream robotics.

Pretraining: Architectures and Datasets. Video diffusion architectures have recently been scaled to increasingly large human video datasets for LfV [306, 67, 68]. Autoregressive transformer architectures have also been employed [105, 289], including on large-scale internet (video-game) data [40]. Bruce et al. [40] and Hu et al. [105] find their video prediction models to scale well with increased compute and model size. Other works [232, 180, 290] use the recurrent state-space model of Dreamer [94], a state-of-the-art model-based RL architecture.

Pretraining: Action-conditioning. To add action-conditioning information to the video predictor, Yang et al. [306] and Sohn et al. [251] pretrain jointly on both video and robot data (the robot data contains action labels used for conditioning). Otherwise, other works have pretrained an alternative-action video predictor $p_{\text{alt}}(o_{t+1}|o_t, \hat{a}_t)$ [306, 40]. Conditioning on an $\hat{\mathcal{A}}$ allows for more control over video generations, and often be a mapping can be obtained from $\hat{\mathcal{A}}$ to the robot action-space [221, 227]. LfV video predictors have been trained to be conditioned on: language-as-actions [306, 68]; single-step latent actions [221, 40]; goal images [67]; future grasp-location and post-grasp waypoint affordance information [180]; or motion information [319, 279]. Yang et al. [306] condition on several different action-spaces, including robot actions, language actions, and camera motions.

Downstream: Adapting the Video Predictor. A video prediction model can be adapted via finetuning on D_{robot} [67, 7]. However, few works directly finetune an action-free video predictor $p(o_{t+1}|o_t)$ into an action-conditioned dynamics model $p(o_{t+1}|o_t, a_t)$. Mendonca et al. [180] pretrain on video with affordance-based conditioning, and add optional robot action-conditioning during finetuning. Seo et al. [232] find naive finetuning on robot data to result in the erasure of pretraining knowledge, and instead 'stack' an action-conditioned model on top of the pretrained model. In related work, Yang et al. [305] demonstrate that the score function of a large pretrained video diffusion model can guide the generations of a smaller task-specific video model. Adaptation can involve obtaining a mapping from $\hat{\mathcal{A}}$ to the robot action space [221] (Section 4.3.1 elaborates on methods that map from $\hat{\mathcal{A}}$ to $\mathcal{A}$).

Downstream: Using the dynamics model. Once the dynamics model has been obtained it has been used in several ways in the LfV literature. (1) *As a simulator*: $p(o_{t+1}|o_t, a_t)$ can be used to generate synthetic data to help train a policy [306, 40]. (2) *As a differentiable simulator*: similarly, one can backpropagate through model-generated rollouts to help train a policy [232, 290, 94]. (3) *For planning*: the dynamics model can be used for standard model predictive control [221, 180] or tree-search [68].

These use-cases often require evaluation of the model-generated trajectories. Some works learn reward/value estimates via a downstream finetuning stage, making use of reward-labelled robot data [232]. Others use video data to help learn a reward function [180, 221, 306, 68]. See Sections 4.3.3 and 4.3.4 for details regarding methods for learning reward and value functions from video.

Discussion. Performance gains: Bruce et al. [40] show their video-game foundation model to exhibit initial signs of impressive generalization. When pretraining with large human video datasets, several works have demonstrated moderate but significant performance gains [180, 290]. Yang et al. [306] and Du et al. [68] demonstrate that large-scale pretrained video diffusion models can be particularly useful in long-horizon tasks.

Pros and cons: These results, combined with the possibility that dynamics models can learn from non-expert and off-policy internet data more easily than policies [315], make LfV dynamics model approaches appear promising. However, we note that, like policies, certain approaches may suffer from missing low-level information in video.

Future directions: (1) An obvious route to improved performances here is to continue to scale our video dataset and model sizes [40, 105]. Experimenting with scalable alternative-action techniques [40] will be useful for conditioning foundational video predictors. (2) A hierarchical world/dynamics model [144], where the higher levels are learned from video data and the lower-levels are learned from robot data, is another direction to explore. (3) An underexplored direction is to combine standard analytic simulation [267, 175] with video prediction simulation [306] to obtain the benefits of both. (4) Following synthetic data trends in other fields [287, 150], future work may seek to use foundational video predictors to generate synthetic video rollouts [306, 40], or augment and improve the coverage of existing data via video editing [316].

4.3.3 *Reward Functions.*

The reward function is an essential component of an RL algorithm. However, manual reward design is difficult to scale to complex and unstructured real-world generalist robot settings. LfV research has sought to tackle this issue by extracting visual reward functions from video data. In this section, we detail any method that uses video data to help relabel transition tuples (o_t, a_t, o_{t+1}) to (o_t, a_t, r_t, o_{t+1}), thus allowing the tuples to be used for online or offline RL.

Extracting reward functions from video. We now describe the main clusters of methods for extracting and constructing reward functions from video data.

- *Video-language Model Rewards.* These methods specify the task via language and use a video-language model (VidLM) (see Section 4.1) to provide a reward signal. We identify two categories of methods here. *Video-text similarity*: A dual encoder model can be trained to embed videos and language into the same representation-space [297, 327]. From such a model, a reward can be defined as the similarity (in embedding space) between a language task-description and a video of the robot's behaviour [77, 65, 253]. Baumli et al. [22] study dual-encoder image-text rewards in detail. *Visual question answering (VQA)*: Reward information can be obtained from a video-to-text model (see Section 4.1.3) by asking it whether a task has been completed in a video. A dense reward could be obtained by asking the model to score the robots progress, or by using the model to provide feedback within an 'RL-AI-F' framework [133]. VQA rewards have been used with images as input [69, 304]. Liu et al. [166] employ video-based VQA rewards, but the limited capabilities of current video-to-text models has hindered progress here.
- *Video-predictors as Reward Functions.* A video prediction model $p(o_{t+1}|o_t)$ can be converted into a reward function [335, 74, 110]. These methods define a reward based on the likelihood of the robot video under the video predictor. This encourages the robot to match the behaviour distribution of the video data. Language-conditioned video prediction may allow these approaches to scale to diverse, non-expert internet video [74].
- *Representational Similarity to a reference.* These methods define the reward as the similarity between representations of the robot's observation(s) and a reference observation (i.e., a goal image or demonstration video). We outline two distinct approaches here. *Standard deep representations:* 'Standard' learning objectives on video data can provide representations useful for measuring the similarity between the current observation and a goal image. Ma et al. [171] use a time-contrastive objective, which gives implicit measurements of the temporal distance between two images. Hu et al. [108] find that many standard deep representations can be effective, though masked autoencoding-based representations perform poorly. *Representations that address LfV distribution shift:* Distribution-shifts (such as embodiment differences)

can prevent meaningful comparison between human and robot videos. Thus, previous work has used representations designed to explicitly ignore such distributions shifts (see Section 4.2.2) when defining their LfV reward. This has included the use of: embodiment-invariant representations [321]; object-centric representations [140, 245, 46]; the retargeting of poses from human to robot embodiments [176, 211]; and viewpoint invariant representations [234, 12].

- *Potential-based shaping with value functions.* A value function, $V(o_t)$, pretrained from video can provide a dense reward via 'potential-based shaping': $r_t = V(o_{t+1}) - V(o_t)$. In the LfV literature, such rewards have been defined using value functions pretrained on video data via TD learning [45], or time-contrastive learning [171].
- *Generative adversarial imitation.* Generative adversarial imitation learning [98] has been used to encourage the robot to match the behaviour distribution of a video dataset during online learning [271]. This has often required the use of representations that explicitly ignore LfV distribution shifts [254, 211].
- *Other methods.* A task classifier can be trained on task-labelled video data to provide downstream rewards [50, 242]. Several methods use the number of steps-to-completion in the video as a proxy label to train a video reward model [306, 71]. Others encourage similarity to a video-obtained behavioural prior: such as a video-pretrained policy [311], or a human affordance distribution [15].

Downstream Usage: Transfer Mechanisms. A video-pretrained reward function can be used zero-shot in the downstream robot domain [74]. Other works further finetune the reward function on in-domain robot data [253]. The LfV reward could be used as the sole reward for the robot, or it could be used as an exploration or shaping bonus, in addition to a sparse task reward [311, 45]. Adeniji et al. [3] pretrain the policy using the LfV reward, before finetuning it on a manually defined task reward.

Discussion. Pros and cons: Learning accurate reward functions purely from video data may be more feasible than other RL KMs. Policies or dynamics models may need access to non-visual information (e.g., forces) in the downstream robot domain, whereas reward functions can often operate solely using visual information and thus can be used zero-shot after video pretraining [74]. However, targeting other KMs via LfV (i.e., policies and dynamics models) may better reduce demands on the robot data and better aid generalization beyond D_{robot}.

Future directions: (1) The most promising approaches are likely those that can leverage pretrained video foundation models: i.e., video-language model rewards and video-predictors as reward functions. (2) One avenue is to combine LfV reward functions with other LfV KMs. For example, finetuning an LfV policy via online RL and LfV rewards. (3) An underexplored direction is to use LfV rewards to augment reward labels in offline RL data. (4) LfV reward functions may prove useful for detecting safety-related metrics when deploying robots in the real world [90]. (5) Many LfV reward functions are differentiable, making them suitable for use in certain model-based RL algorithms [94, 122]. (6) Shaped reward functions could be extracted from video-to-text models via RL from AI feedback (RL-AI-F) methods [133].

4.3.4 Value Functions. Value functions are an essential component of most deep RL algorithms [231, 93]. A small but distinct line of LfV research pretrains models that closely resemble value functions using video data. We outline methods that do so below.

Pretraining: TD-learning. The temporal-difference (TD) learning objective is commonly used to learn value functions in RL [257]. However, video data is missing important action, reward, and goal labels that are often required during TD-learning.

- *Missing action labels:* Action labels are required to obtain a state-action value function $V(o_t, a_t)$, which can be more useful than a 'state' value function $V(o_t)$ [257]. Thus, LfV research has sought to train value functions conditioned on 'alternative-actions' (see Section 4.2.1): $V(o_t, \hat{a}_t)$. This has been achieved by conditioning on: sub-goals [27, 82]; next observations [73]; single-step latent actions [45]; and IDM-obtained pseudo-actions [46].

- *Missing reward and goal labels:* TD-learning requires reward labels. Meanwhile, TD-learning of multi-task/goal value functions requires task/goal labels. To obtain reward labels for TD learning, Bhateja et al. [27], Ghosh et al. [82], and Park et al. [199] use hindsight goal relabelling: a sparse reward is defined as $r = (o == g)$ (where o is the current observation and g is the goal observation). Chang et al. [46] leverage object labels and off-the-shelf object detectors to provide goal and reward labels in navigational video data. Other works assume a single task setting [73], or assume access to reward labels in the video data [73, 45]. These are not scalable assumptions. The LfV reward functions from Section 4.3.3 could be applicable here, while video-to-text models (see Section 4.1.3) could provide goal labels in textual form.

Pretraining: Time-contrastive learning. Time-contrastive objectives induce a temporally smooth representation, and a value function can be defined by measuring the distance between the current observation and a goal image in the representation space [171]. Importantly, these objectives do not require action labels for video pretraining. Quasimetric functions [283] and temporal cycle-consistency objectives [321] have been used to learn similar representations. Note, the requirement for goal images is a limitation of these approaches.

Pretraining: Others. Edwards and Isbell [71] regress a value function towards a heuristic value; the number of timesteps remaining in the video. This approach assumes expert behaviour in the video. Du et al. [68] take a similar approach, finetuning a VLM to give the heuristic value estimates. Liu et al. [165] use critical state identification to aid value predictions, but this assumes access to reward labels in the video data.

Downstream Usage. We now briefly outline how video-pretrained value functions have been used in the downstream robot domain in the literature. (1) *As a value function:* A video-pretrained value function can be directly used downstream in a standard fashion if: it is an action-conditioned $V(o_t, a_t)$ [46]; or it is an alternative-action-conditioned $V_{\text{alt}}(o_t, \hat{a}_t)$ and a mapping from $\hat{a}$ to a can be obtained; or an action-conditioned dynamics model $p(o_{t+1}|o_t, a_t)$ is available, allowing an unconditioned $V(o)$ to be used for tree-search planning [68, 45]. (2) *Representation transfer:* Bhateja et al. [27] initialise their downstream value function and policy representations from a video-pretrained value function. (3) *Potential-based reward shaping:* A reward function can be defined as: $r = V(o_{t+1}) - V(o_t)$, and be used for downstream online RL [71, 311, 45]. It may be desirable to do this if: the pretrained value function is not fully reliable but can provide useful auxiliary rewards to guide exploration; or the value function is not action-conditioned, so cannot be used as a $V(o_t, a_t)$ for Q-learning [257]. (4) *TD bootstrapping:* A $V(o_t)$ can still be used to accelerate downstream TD-learning of a state-action value function $V(o_t, a_t)$ by using its estimates for the bootstrap term in the bellman backup [71].

Discussion. Research into learning value functions from video has been relatively scarce. This is perhaps due to several associated challenges. First, as noted above, there are issues related to missing action, reward, and goal labels in video. Second, value functions estimate returns under a particular policy, but the behaviour in video data is highly multi-modal. Third, TD learning from video may run into common issues seen in the offline RL literature [146], though the offline RL literature does present potential solutions [138, 334]. We also note that, if the goal is to obtain a policy, it may be easier to attempt this directly from video (see Section 4.3.1). Nevertheless, Bhateja et al. [27] show that TD-learning of value functions from large-scale human video is a promising direction for real-world robotics.

5 Datasets and Benchmarks

This section first provides details and discussions regarding video datasets in LfV (Section 5.1), before moving to the benchmarking of LfV methods (Section 5.2).

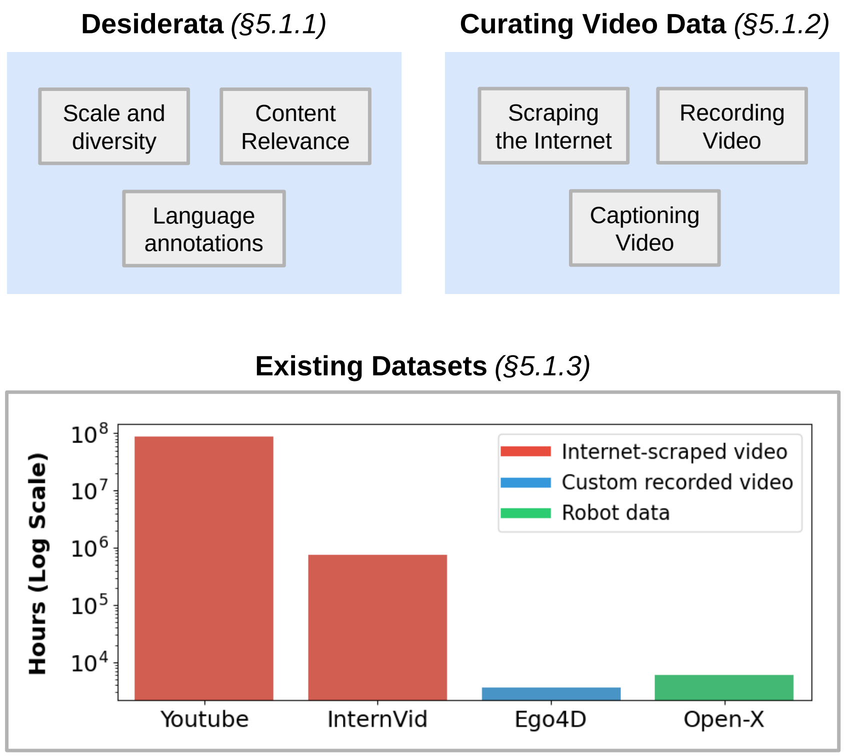


Fig. 8. **LfV Datasets** (Section 5.1). The log-scale plot at the bottom compares the sizes of the largest curated open-source datasets in three different categories. InternVid [284], Ego4D [87], and Open X-Embodiment [195] are the largest internet-scraped video, custom-recorded video, and robot datasets, respectively (to the best of our knowledge). We estimated the hours in the Open X-Embodiment dataset by assuming the average trajectory length is 10 seconds (at the time of writing, the dataset contained ~2 million trajectories). The number of hours of video on YouTube is a rough calculation [249].

5.1 Datasets

We now turn our attention to the video datasets themselves. In this section (see Figure 8), we will discuss the desired properties of video datasets (Section 5.1.1), summarise methods for curating video data (Section 5.1.2), and review existing datasets and their limitations (Section 5.1.3).

5.1.1 Desiderata. We now discuss the key properties and characteristics desired from our LfV video datasets.

Scale and Diversity. The dataset should be large in scale and high in diversity. Scale can be measured in terms of total video duration. Diversity refers to variety in the content of the data. Increased scale and diversity of training data can reliably improve deep learning performances and generalization [39]. LfV is promising precisely due to the massive scale and diversity of video data available on the internet.

Content Relevance. The content and information in the video data should be relevant to downstream generalist robot settings [129]. For example, the videos should include information regarding the dynamics of the world and how embodied agents can complete physical tasks. Crucially, the video data should have good coverage over the tasks and environments the generalist robot is likely to encounter.

Language Annotations. Language annotations can be useful for training language-conditioned models [24, 38] and to aid semantic representation learning from video [327]. High-level annotations can be useful for learning abstracted representations. Granular annotations—for example, detailed descriptions of actions, spatial information, and object relations—can aid the learning of low-level features that are particularly relevant to robotics.

Other Desiderata. (1) Many LfV methods assume continuity within a clip (i.e., an absence of sharp scene transitions) [40, 306]. (2) The use of high-resolution video may benefit finer-grained downstream behaviours [152]. (3) Longer video clips may benefit long-horizon representations and memory useful in multi-step tasks [163, 278]. (4) Video can be paired with other useful information, besides language annotations. This includes object bounding boxes, object segmentations, or human-pose estimates — these labels can aid representation learning efforts [15], serve as action labels [103], or inform automated language captioning [31]. Modalities such as audio or depth information [87] can provide additional information beyond what is contained in RGB pixels.

5.1.2 Curating Video Data. Here, we give concrete details on techniques for curating video datasets. We discuss techniques for scraping video from the internet, techniques for recording custom video, and techniques for (manually and automatically) annotating video.

Scraping Video from the Internet. Diverse, large-scale video datasets can be obtained from internet repositories of pre-existing video data. Techniques here focus on ensuring only relevant and high-quality videos are scraped from these repositories. A typical pipeline for scraping internet video involves: (1) *Formulating a pool of query prompts* used to search the repository for candidate video with relevant content. Previous works have constructed this pool, for example, via: surveys of human time-use [87, 284]; crowd-sourcing [235, 85]; or by using LLMs to parse action prompts from text corpora [284] (2) *Post-processing the pool of candidate videos.* This has included processing videos into clips which contain no 'cuts' between scenes [33] and using video meta data (e.g., view counts) to identify and remove low-quality video [300, 181, 187]. Automated metrics such as optical flow [33], image-language model embeddings [230, 33], and video 'realism' estimates [239] have also been used to filter video data. (3) *Ensuring adherence to legal, ethical, and safety standards.* This should include only selecting appropriately licensed data [250], ensuring privacy by anonymizing faces [250], filtering out unsafe content [121], and minimizing and explicitly acknowledging potential biases [250, 181]. (4) *Re-annotating the videos to improve the quality of labels.* We describe video captioning methods in detail in the below paragraphs.

Recording New Video. Manually recording custom videos can be an effective (but expensive) means of collecting video relevant to specific robot tasks [235, 86] or embodiments [235, 237]. Large-scale recording of custom video datasets has mainly been achieved via crowd-sourcing [85, 87]. Damen et al. [59] and Grauman et al. [87] ask participants to record their daily lives, while Goyal et al. [85], Sermanet et al. [235], and Grauman et al. [86] provide participants with text instructions regarding the behaviours they should perform. Video can be collected with a diversity of embodiments, including human arms, human-controlled manual grippers, and robot arms [235, 237]. Recording first-person video [87]—video which may be less common on the internet—could be particularly relevant to a robot with first-person observations. Attempts to record custom video should adhere to ethical standards — including minimizing bias by ensuring diversity across various demographic and contextual dimensions, and ensuring individuals in the video give consent or have their privacy protected (e.g., via face blurring) [86].

Manual Captioning. The benefits that language captioning of video can provide are outlined in Section 5.1.1. Previous works have often obtained descriptions of the contents of video data via manual human labelling [33, 59, 86, 87]. This can be performed by third-party workers or (where applicable) the person recording the video. Versus to typed annotations, Damen et al. [59] find spoken annotations to result in higher quantity and quality of captions. Grauman et al. [86] have the person recording the video explain their thought processes whilst performing actions, and then have third-party annotators provide retrospective expert commentaries. After initial annotation is performed, Maaz et al. [172] prompt workers to augment the descriptiveness of the video captions.

Automated Captioning. Manual human annotations are expensive [87]. For larger-scale datasets, automated annotation pipelines are promising. Various methods have been proposed here.

- *Automatic speech recognition (ASR), metadata, and others.* ASR converts speech in the video's audio into text [181, 300, 322]. However, raw ASR captions can be noisy and unrelated to the contents of the video. Another source of information is corresponding metadata; such as descriptions, tags, and titles [284]. Bain et al. [17] obtain captions from the alt-text HTML attribute associated with web images and videos [17] ('Alt-text' in Table 1). Nagrani et al. [187] start with a dataset of high-quality image-caption pairs, then mine video clips with similar frames and transfer the captions to those clips ('Transfer' in Table 1).
- *Off-the-shelf vision models.* Video-to-text captioning models are an option [33], but poor captioning capabilities have limited previous usage. Alternatively, more reliable image-captioning [33, 172] or object-detection [323, 172, 215, 31] can be employed to gather information regarding the content in a video. The methods in this bullet are referred to as 'Generated' in Table 1.
- *LLM processing.* Blattmann et al. [33], Maaz et al. [172], and Wang et al. [284] use LLMs to summarise keyframe captions and full-clip captions into a single video caption. LLM post-processing can be used to filter out inconsistent captions across sources or frames [172]. Combining multiple sources of information via LLM processing may result in more detailed and accurate captions.

5.1.3 Existing Datasets. In this section, we provide an overview of existing video datasets relevant to LfV. We aim to highlight datasets that satisfy key desiderata from Section 5.1.1, and thus are promising for training foundational video and/or robotics models.

Overview of Existing Datasets. Table 1 presents details of a representative set of existing video datasets. We now discuss these datasets in more detail.

- *Large-scale, internet-scraped datasets.* These can span up to several decades worth of video data [284, 300]. They are diverse, capturing human behaviours in tasks and environments sampled from a global population. These datasets have been constructed via: query-based searches of YouTube [284, 181, 255]; aggregating video from previous datasets [322]; or scraping video from various webpages [17]. Language captions of these large datasets are usually obtained via automated methods. These datasets have commonly been used in initial attempts at training video foundation models [327, 280] (see Section 4.1).
- *Manually collected datasets.* These are generally not as large or diverse as the largest web-obtained video datasets. However, many contain content highly relevant to robotics, and these datasets have often been used in past LfV research [188, 306, 289]. For example, Ego4D [87] contains egocentric video of participants going about their daily lives, and RoboVQA [235] contains video of teleoperated robots and humans performing long-horizon household tasks. Here, it is common for captions to be provided manually by crowd-sourced workers.
- *Annotations.* All of the datasets in Table 1 come with text annotations. Other annotation modalities are sometimes provided. Grauman et al. [87, 86] occasionally collect corresponding audio, 3D meshes, eye-gazes, stereo, and synchronized video from different viewpoints. Shan et al. [239] provide human-hand labels for 100k images in their video dataset. Note, internet-obtained video datasets generally lack these other annotations as they may be costly to add manually, or may require customized video recording setups.

[1] These datasets are not publicly available at the time of writing.

Table 1. **Existing video datasets**. Listed are: (top) large-scale, internet-scraped video datasets, and (bottom) robotics-relevant, manually-recorded video datasets. The datasets are ordered by decreasing total video duration. Details regarding 'Caption Type' can be found in Section 5.1.2.

Dataset	Content	Size (hours)	# Clips	Caption Type	Collection Method
InternVid [284]	YouTube	760,000	230M	Generated	Internet
HD-VILA-100M [300]	YouTube	370,000	103M	ASR	Internet
YT-Temporal-180M [322]	YouTube	-	180M	ASR	Internet
WTS-70M [255]	YouTube	190,000	70M	Metadata	Internet
HowTo100M [181]	Instruction	134,000	136M	ASR	Internet
WebVid-10M [17][1]	YouTube	52,000	10M	Alt-text	Internet
VideoCC3M [187][1]	YouTube	18,000	6M	Transfer	Internet
100 Days of Hands [239]	Actions	3,100	27k	Metadata	Internet
Ego-4D [87]	Everyday	3,600	28k	Manual	Manual
Ego-Exo-4D [86]	Skilled	1,400	6k	Manual	Manual
SS-v2 [85]	Actions	245	221k	Manual	Manual
RoboVQA [235]	Everyday	230	98k	Manual	Manual
Epic-Kitchens-100 [58]	Cooking	100	700	Manual	Manual

Discussion. The largest internet-curated video datasets are two orders of magnitude larger than the largest manually-recorded video datasets. Yet, these still barely scratch the surface of the full range of video content available online (see Figure 8). As such, we advocate for continued efforts into curating ever-larger internet video datasets for LfV (whilst upholding high legal, ethical, and safety standards). Crucially, new curation efforts should optimise for the full range of desiderata discussed in Section 5.1.1). Low-quality language annotations are a key limitation of current internet-curated video datasets. This can be addressed via improved automated captioning methods. Finally, we note that efforts to curate smaller-scale, higher-quality video datasets can still provide value. For example, such high-quality datasets can be used for finetuning models after pretraining on larger, lower-quality data [33].

5.2 Benchmarks

Benchmarks can be crucial for catalysing rapid research progress in a given area [63]. In this section, we first give recommendations for how an LfV benchmark should be designed (Section 5.2.1). We subsequently review relevant benchmarks from the literature, commenting on limitations and proposing improvements (Section 5.2.2).

5.2.1 Designing LfV Benchmarks. In the paragraphs below, we first give recommendations regarding the metrics an LfV benchmark should measure, before outlining different possible categories of LfV benchmarks and their corresponding desiderata.

Metrics. An LfV benchmark should serve to evaluate either the capabilities of a policy obtained via an LfV approach, or the effectiveness of an LfV algorithm at producing a policy, under certain constraints (e.g., when constrained to use a fixed dataset). More specifically, the benchmark can evaluate the extent to which the potential benefits of LfV (Section 3.3) have been obtained. This can be achieved via the following quantitative metrics:

(1) *Performance in-distribution of $\mathcal{D}_{robot}$.* Performance (e.g., task success rate) of the robot in settings in-distribution of $\mathcal{D}_{robot}$, after training on $\mathcal{D}_{video}$ and a fixed $\mathcal{D}_{robot}$.

(2) *Data-efficiency in-distribution of* $\mathcal{D}_{robot}$. The quantity of data in $\mathcal{D}_{robot}$ required to reach a certain performance level in settings in-distribution of $\mathcal{D}_{robot}$, when also training on $\mathcal{D}_{video}$.
(3) *Generalization beyond* $\mathcal{D}_{robot}$. The performance of the robot in settings out-of-distribution of $\mathcal{D}_{robot}$, after training on $\mathcal{D}_{video}$ and a fixed $\mathcal{D}_{robot}$.

These metrics cover the benefits of 'generalization beyond $\mathcal{D}_{robot}$' and 'improvements in-distribution of $\mathcal{D}_{robot}$', as outlined in Section 3.3. However, the benefit of 'emergent capabilities' likely must be evaluated qualitatively.

Finally, the benchmark should also inform us on how scalable the LfV method is to diverse internet data and the generalist robot setting, and how well it can handle fundamental LfV challenges (Section 3.4). Specific design choices that can be made here are discussed in later paragraphs.

Categories. All LfV benchmarks should include a fixed set of evaluation environments, $\mathcal{M}_{\text{eval}}$, in which the LfV policy, π_{lfv}, is to be evaluated. However, LfV benchmarks can differ based on whether they specify the datasets $\mathcal{D}_{\text{video}}$ and $\mathcal{D}_{\text{robot}}$ that are to be used to train the policy. Concretely, we identify the following possible categories of LfV benchmark:

- $B_{\mathbf{e}} = \{\mathcal{M}_{\mathbf{eval}}\}$**.** A π_{lfv} trained on any $\mathcal{D}_{\text{video}}$ and any $\mathcal{D}_{\text{robot}}$ is evaluated on a fixed $\mathcal{M}_{\text{eval}}$.
- $B_{\mathbf{e\text{-}r}} = \{\mathcal{M}_{\mathbf{eval}}, \mathcal{D}_{\mathbf{robot}}\}$**.** A π_{lfv} trained on any $\mathcal{D}_{\text{video}}$ and a *fixed* $\mathcal{D}_{\text{robot}}$ is evaluated on a fixed $\mathcal{M}_{\text{eval}}$.
- $B_{\mathbf{e\text{-}r\text{-}v}} = \{\mathcal{M}_{\mathbf{eval}}, \mathcal{D}_{\mathbf{robot}}, \mathcal{D}_{\mathbf{video}}\}$**.** A π_{lfv} trained on a *fixed* $\mathcal{D}_{\text{video}}$ and a *fixed* $\mathcal{D}_{\text{robot}}$ is evaluated on a fixed $\mathcal{M}_{\text{eval}}$.

Benchmarks that fix the datasets (e.g., $B_{\text{e-r-v}}$) can provide a fair comparison of LfV algorithms. However, they may end up being 'toyish' in practice. Benchmarks that do not fix the data (i.e., B_{e}) are useful for evaluating the performance of an obtained π_{lfv}.

Desiderata. We now give details regarding the desiderata for each potential component of an LfV benchmark.

- $\mathcal{M}_{\mathbf{eval}}$. The evaluation environments should be analogous to the generalist robot settings we are interested in. (1) *Relevance:* The environments and tasks should resemble those in generalist robot settings. (2) *Diversity* in $\mathcal{M}_{\text{eval}}$ allows us to measure how π_{lfv} will handle diverse and unseen real-world scenarios. (3) *Realism:* The benchmark should include challenges that may be faced in real-world environments (such as noisy observations), and its physics should be sufficiently realistic. To ensuring $\mathcal{M}_{eval}$ bears on the challenge of 'missing low-level information in video', the tasks in $\mathcal{M}_{eval}$ could be setup to require perception of information not contained in video data.
- $\mathcal{D}_{\mathbf{robot}}$. (1) $\mathcal{D}_{\text{robot}}$ should be drawn from a *subset* of the tasks and environments in $\mathcal{M}_{\text{eval}}$, allowing us to measure performance both in and out-of-distribution of $\mathcal{D}_{\text{robot}}$. (2) To ensure LfV generalization (see Figure 2) is indeed possible in the benchmark, the robot dataset should have good coverage over all possible low-level "atomic" actions.
- $\mathcal{D}_{\mathbf{video}}$. Firstly, $\mathcal{D}_{\text{video}}$ should not contain action labels. Other desiderata depend on the extent to which we wish $B_{\text{e-r-v}}$ to be a 'toy' setting for testing LfV algorithms.
  Toy setups can allow for excellent control over which LfV challenges (Section 3.4) are faced. (1) A toy setup can ensure $\mathcal{D}_{\text{video}}$ has good coverage over the environments and tasks in $\mathcal{M}_{\text{eval}}$. (2) Characteristics of $\mathcal{D}_{\text{video}}$ can be controlled to establish distribution shifts between $\mathcal{D}_{\text{video}}$ and $\mathcal{M}_{\text{eval}}$ (e.g., embodiment gaps, viewpoint differences). (3) The challenge of 'controlability' (see Section 3.4) can be toggled by managing the extent to which changes in the video are due to effects beyond a single agent's actions.
  Approximating the scaled-up LfV setting: Here, $\mathcal{D}_{\text{video}}$ should resemble internet video in terms of its scale, diversity, and content. Thus, $\mathcal{D}_{\text{video}}$ should ideally consist of real-world human videos scraped from the internet. Doing so will inherently present several LfV challenges, and allow for evaluation of the scalability of the LfV method.

5.2.2 Existing Benchmarks. This section briefly details existing LfV benchmarks, before discussing future directions.

$B_e = \{\mathcal{M}_{eval}\}$. A number of benchmarks provide an $\mathcal{M}_{\text{eval}}$ relevant to LfV research. (1) *Toy settings:* The diversity of certain complex or open-ended video game environments [77, 142] can provide an excellent setting for testing LfV policy generalization. (2) *Robotics simulators:* Robotics-relevant simulated environments include benchmarks focused on low-level motor control and object interaction [179, 314, 89, 162, 141, 175, 209]. (3) *Embodied AI simulators:* Some benchmarks focus on household and everyday mobile manipulation tasks while abstracting away low-level control to focus on higher-level planning [208, 135, 148]. (4) *Real-robot setups:* Real-world robot evaluations are also available [4], but may be more costly and time-consuming than simulated evaluation. To address this, Li et al. [156] design a simulated benchmark specifically for evaluating real-world policies.

$B_{e\text{-}r} = \{\mathcal{M}_{eval}, \mathcal{D}_{robot}\}$. Several benchmarks pair a $\mathcal{D}_{\text{robot}}$ with an $\mathcal{M}_{\text{eval}}$. Mees et al. [179], Lynch et al. [169], and Liu et al. [162] provide demonstration or play data for simulated tabletop manipulation tasks. Gu et al. [89] includes mobile manipulation tasks. Pumacay et al. [209], Xie et al. [294], and Nasiriany et al. [190] provide multiple different tasks and environments (along with corresponding $\mathcal{D}_{\text{robot}}$ generation methods), which may be useful for testing LfV generalization. Lastly, there exist relatively large real-world robot datasets which are not paired with any specific benchmark [195, 129].

$B_{e\text{-}r\text{-}v} = \{\mathcal{M}_{eval}, \mathcal{D}_{robot}, \mathcal{D}_{video}\}$. To the best of our knowledge, there are few benchmarks that specify a fixed $\mathcal{D}_{\text{video}}$ and a fixed $\mathcal{D}_{\text{robot}}$. Fan et al. [77] provide a $\mathcal{D}_{\text{video}}$ containing 730k YouTube videos in a Minecraft setting. Many previous LfV works have constructed a $\mathcal{D}_{video}$ by stripping action labels from a $\mathcal{D}_{robot}$ [232, 227, 288]. This can allow for easy setup of a scenario where $\mathcal{D}_{\text{video}}$ has good coverage over $\mathcal{M}_{\text{eval}}$, but can neglect distribution shifts seen between the video data and the robot domain in realistic LfV settings.

Discussion. We provide recommendations for improvements in LfV benchmarking.

- *Improving the diversity in $\mathcal{M}_{eval}$.* The diversity in most suitable robotic simulators [179, 169, 162, 89, 148] is still limited in terms of the tasks, environments, and objects presented. Improving diversity will allow us to better assess the applicability of the LfV method to generalist robot settings. One possibility here is to use procedural generation [62] or LLM-assisted environment design [292]. Another is to aggregate multiple $\mathcal{M}_{\text{eval}}$'s within a common framework [141].
- *Establishing popular $B_{e\text{-}r}$ and $B_{e\text{-}r\text{-}v}$ LfV benchmarks.* Establishing benchmarks in these categories (currently there are no popular options) would provided an improved ability to compare LfV algorithms; past works have often chosen different $\mathcal{D}_{\text{video}}$'s [188, 306, 227] and thus are difficult to compare. Any new LfV benchmark should be designed following the recommendations outlined in Section 5.2.1.

6 Challenges and Opportunities

We now provide a comprehensive discussion of challenges and opportunities for future LfV research, based on our analysis of the existing literature. First, we give high-level recommendations for future LfV research (Section 6.1). Second, we detail promising directions for utilising video foundation models and techniques for LfV (Section 6.2). Third, we highlight approaches for overcoming previously identified LfV challenges (Section 6.3). We conclude by discussing other challenges in generalist robotics may not be solved via scaling to larger datasets (Section 6.4).

6.1 High-level Recommendations

Following our the analysis of the LfV literature in Section 4, we now provide some high-level recommendations for future LfV research.

Focus on scalable approaches. Methods that can scale well to large, diverse internet video may be particularly effective. Many previous LfV works limit scalability by making strong assumptions on the nature of the video data or the downstream robot setting [18, 254]. Instead, we advocate for methods that can use general learning objectives to extract knowledge from heterogenous video data (see the foundation model approaches in Section 4.1). Methods that can learn from unlabelled and suboptimal video data are most scalable in theory [40], though methods that leverage language labels can still be highly effective [38, 306]. Finally, focusing initially on LfV approaches that can learn from offline robot data (or that can perform online RL in simulation) may be sensible, due to the impracticalities of online real-world RL.

Target the key benefits of LfV via policies and dynamics models. Future LfV research should focus on obtaining the most promising LfV benefits (Section 3.3); namely, improving robot data efficiency and improving generalization beyond $\mathcal{D}_{\text{robot}}$. Much past LfV research has sought to extract reward functions from video (Section 4.3.3). Whilst these approaches are useful, learning policies (Section 4.3.1) and dynamics models (Section 4.3.2) is likely most promising for reducing reliance on robot data (as is discussed in their respective sections).

Improve LfV datasets. Improving and scaling video datasets is a reliable way to improve our LfV models. In Section 5.1.3, we recommend scraping larger-scale video data from the internet, and captioning this data with high-quality language annotations. We advocate for the open-sourcing of such curated video datasets.

Combine LfV with alternative approaches, like simulation. Any practical LfV approach should seek to collect as useful a $\mathcal{D}_{\text{robot}}$ as possible. In some cases, video data could aid the collection of (real and simulated) robot data. Automatic generation of diverse and realistic simulated environments [292, 76] could be aided by video data [309] and video models [38]. Here, video foundation models could be used to provide dense reward signals in diverse simulated environments (Section 4.3.3, Li et al. [149]). Elsewhere, we note that the use of video data is not restricted to monolithic foundation model approaches or pure end-to-end learning approaches. Components in modular approaches, such as specialist agents [303] or vision systems, may benefit from the use of internet video data. In hybrid approaches, where deep learning is combined with more classical or structured control methods [57, 111, 182], video data can benefit the deep learning component.

Improve evaluation of LfV methods. Research should explicitly measure the extent to which a method can handle LfV challenges and can provide the potential benefits of LfV. In Section 5.2.1, we outline concrete metrics for this. It is often difficult to identify these metrics in existing LfV research. In Section 5.2.2, we advocate for the design of improved LfV benchmarks that can offer these metrics off-the-shelf.

6.2 Video Foundation Models for LfV

A promising LfV direction is to utilise large-scale internet video data to help train large foundation models (FMs) for robotics. Here, we outline related research avenues and key points of discussion.

Improving Video FMs. Advancing video FMs and techniques may be a key driver of progress for scalable LfV approaches (Section 4.1). Key issues to tackle here include addressing dataset limitations and prohibitive computational requirements. More details on avenues for improving video FMs are discussed in Section 4.1.4.

Leveraging Pretrained Video FMs. Generic video foundation models can be adapted into RL knowledge modalities to aid robot learning. Strong baselines here will include finetuning the video FM into an action-outputting policy using offline robot data [37, 289], or (if the video FM has video generation capabilities) finetuning it into a robot dynamics model. Data-efficient and computationally efficient adaptation mechanisms [106, 305] may be useful here.

Customising Video FMs for LfV.. The most effective LfV approaches will likely use pipelines fully optimized for robotics. We now touch on directions for customising video FMs and techniques for LfV.

- *Improve the capabilities of generic video FMs in areas relevant to robotics.* We could, for example, improve the physics realism of generic video prediction models (e.g., via RL finetuning; see Black et al. [29]), or the fine-grained understandings of generic video-to-text models (e.g., via improved fine-grained language captioning of the video data). The inclusion of domain-specific robot videos in generic video FM pretraining could improve its performance in the downstream robot setting.
- *Pretrain control-centric video FMs on $\mathcal{D}_{video}$.* There are various options for obtaining models that are more 'control-centric' from video data. The use of action representations (see Section 4.2.1) can add explicit action information to the video FM [40, 227]. The model could be trained using control-centric objectives, such as TD-learning [27]. Video data could be paired with robotics-relevant low-level information (where available); e.g., 3D depth information can improve crucial 3D understandings [331, 318], or frozen video features could be combined with features from different modalities [161].
- *Pretrain control-centric FMs on $\mathcal{D}_{video}$ and $\mathcal{D}_{robot}$.* Any-to-any sequence models can be pretrained jointly on video and robot data, and can subsequently act as policies, dynamics models, and high-level planners [251]. Video prediction models can be pretrained on video and robot data, allowing them to be optionally conditioned on robot actions [306]. Action labels in robot data could be used in addition to action representations in video data. It is not yet clear whether pretraining on both $\mathcal{D}_{\text{video}}$ and $\mathcal{D}_{\text{robot}}$ is preferable to pretraining on $\mathcal{D}_{\text{video}}$ before finetuning on $\mathcal{D}_{\text{robot}}$. Future research should clarify this.

Monolithic Vs Compositional Models. Any-to-any sequence modelling is an example monolithic LfV approach [251]. The use of a video generation model to hierarchically condition an action-prediction model is an example compositional LfV approach [7]. Versus monolithic approaches, Du and Kaelbling [66] argue that compositional approaches are more computationally efficient, more data-efficient, and can obtain improved generalization [333]. These advantages are very relevant to the LfV setting. Nevertheless, monolithic models may benefit from positive transfer between the different data modalities and tasks they handle, and can be optimized end-to-end. Future research should seek to better compare monolithic and compositional LfV approaches.

Open-sourcing Video FMs. We advocate for increased open-sourcing of video foundation models. Open-sourced models will make cutting-edge LfV research more accessible to academic researchers, accelerating progress in LfV [131].

6.3 Overcoming LfV Challenges

In Section 3.4, we outlined the key challenges in LfV. Whilst recent LfV research has partially addressed some of these, here we discuss further promising directions.

Bridging the gap to low-level robot information. A key LfV challenge is to minimise the quantity of robot data required, despite the missing low-level information (e.g., forces and tactile information) in the video data (see Figure 2). While it is possible that further scaling the robot and video datasets will implicitly solve this issue, here we discuss more explicit solutions.

- *Incorporate low-level information into video pretraining.* As discussed in Section 6.2, low-level information could be introduced during video pretraining via: the use of video paired with low-level information (e.g., depth info); or the use of robot data during pretraining. Simulated robot data could be cheap source of low-level information.

- *Bypass the need to incorporate low-level information into video pretraining.* A hierarchical approach can mitigate the need for a video-pretrained model to make use of non-visual information. For example, a high-level policy trained on $\mathcal{D}_{\text{video}}$ can propose alternative-actions, and a low-level policy trained on $\mathcal{D}_{\text{robot}}$ can decode these to robot actions [67, 227, 288] (see Section 4.3.1). The use of a coarse-action space (e.g., cartesian control for manipulation; see Brohan et al. [36]) may achieve a similar effect. One possibility to collect robot data with the hierarchical/coarse action-space policy, and subsequently finetune for finer-grained control.

Recovering action information from video. The missing action label problem in LfV can be partially addressed using alternative-action representations (see Section 4.2.1). We recommend future research to: (1) Scale existing methods to realistic and diverse internet video. (2) Explicitly compare the pros and cons of different action representations, and measure the extent to which each can help obtain the benefits of LfV (see Section 3.3). (3) Develop improved, novel action representations. Perhaps by combining the benefits of existing options [310].

Tackling distribution shift. Previous LfV works have attempted to address LfV distribution shifts using specific inductive biases (see Section 4.2.2). However, these have not scaled to heterogenous internet video data. We advocate for implicitly addressing distribution shifts by scaling flexible methods to large, diverse internet video data, and, when possible, additionally training on (real and simulated) robot data. Compositional approaches are another option for obtaining generalization under distribution shift [66]. Nevertheless, addressing LfV distribution shift is an open problem.

Mitigating computational demands. The high computational demands of training on video data can be mitigated via improved efficient architectures, taking inspiration from relevant LLM research [280, 164, 19, 118]. The ever decreasing cost of compute will also help here [183]. Improved methods for learning latent spaces in which the video model can operate (versus operating directly in pixel space) can aid computational efficiency in addition to mitigating issues related to noise and redundancy in video [21, 302].

6.4 Is Scaling Enough?

Observing results in other machine learning domains [2, 24, 306, 256], it seems likely that scaling current robot learning approaches to larger datasets will yield improvements in robot capabilities. This survey advocates for the use of internet video data to help us achieve this scaling. However, scaling may be insufficient to take us all the way to generalist robots [139]. Indeed, there may be unique challenges in robotics that do not exist in other domains where scaling has proven successful [184]. In this section, we note challenges that may face LfV approaches that rely on scaling the data under the current paradigm. Instead, solving these challenges may require algorithmic advances.

- *LfV is also reliant on robot data.* LfV approaches augment a robot dataset with (internet) video data. However, any LfV approach will still be bottlenecked by the quantity, quality, and coverage of its robot data. The difficulty of collecting robot data with sufficient coverage over generalist robot settings may be the main barrier to scaling up robotics as seen in other domains, even for LfV approaches.
- *Safety and reliability.* Safety is paramount when deploying robots in the real world — failures can lead to damage to property, costly operational disruptions, or physical harm to people [143]. As such, physical robots may require higher success rates than are required in other generative AI tasks. Unfortunately, deep learning models are known to be brittle and often generalise poorly to unseen scenarios [84, 96]. Scaling to internet data may provide a robot model with improved 'common-sense', improving its reliability and generalization [330, 37]. However, this may not fundamentally solve the problem — it may only improve generalization within the distribution of the internet data, but not beyond it.

- *Long-horizon tasks and memory.* Generalist robots may need to operate over long time horizons. This presents two challenges. First, longer-horizon tasks may necessitate improved planning, reasoning, and subtask success rates. Second, longer-horizon tasks require improved memory capabilities. Current solutions to long-term memory involve scaling up the context-length of the model [163], or storing and retrieving from simple memory databases [198]. It is unclear whether these solutions will be sufficient to fundamentally address the problem.
- *Latency.* Large foundation models currently have long inference times that limit their ability to perform real-time high-frequency robot control [217, 36]. Techniques to reduce inference latency will be beneficial here [118, 252].
- *Continual learning and adaptation.* A generalist robot may regularly encounter novel scenarios. It must be able to adapt to these scenarios appropriately. Scaling can improve generalization and in-context 'meta-learning' abilities [39]. Nevertheless, obtaining true continual learning [1] abilities is still an open problem, and may require further algorithmic advances.
- *Reasoning*. There exists both speculation and empirical evidence regarding the inability of deep learning methods to perform true 'reasoning' [109, 11]. Such reasoning abilities may be crucial for a generalist robot. If deep learning is fundamentally limited in this regard, then scaling to internet video data may be insufficient to achieve general-purpose robots.

Nevertheless, 'Is scaling enough?' is ultimately an empirical question. Given deep learning will likely be a core component of generalist robot research for the foreseeable future, there is value in work pushing scaling-based approaches as far as possible. This will allow us to empirically answer this question and find fundamental and practical limitations that do indeed exist. Identified limitations can be tackled in a targeted manner by combining deep learning with alternative approaches (Section 2).

7 Conclusion

Developing general-purpose robots is a grand challenge in robotics. However, current robot learning approaches are bottlenecked by a lack of robot data. Learning from Video (LfV) methods seek to alleviate this issue by augmenting their training dataset with video data. These methods are promising as internet video comes in vast quantities and contains information highly relevant a general-purpose robot.

In this survey, we have outlined fundamental LfV concepts and conducted a comprehensive review of the LfV literature. We emphasised methods with the potential to scale to large, heterogeneous internet video datasets: following the success of scaling deep learning to internet data in other domains [193, 24], these LfV methods are promising for improving the generality of our robots.

The key takeaways of this survey are summarised below.

- *Fundamental LfV concepts.* We explicitly establish the benefits that can be obtained from LfV (Section 3.3), and key LfV challenges that stand in the way of these benefits (Section 3.4). These fundamental concepts should guide future LfV research agendas.
- *Robot foundation models from internet video.* Video foundation model techniques are promising for extracting knowledge from large, heterogeneous internet video datasets (Section 4.1). In Section 6.2, we discussed directions for leveraging video foundation model techniques for robotics. Developing models customized for robotics is a particularly promising direction. Here, recent monolithic any-to-any sequence models offer a clear path forward. It will also be beneficial to pursue compositional approaches in parallel.
- *Scalable approaches for learning policies and dynamics models.* Our main analysis of the LfV-for-robotics literature (Section 4.3) yielded important takeaways. First, future LfV research should avoid inductive biases that limit scalability (i.e., those seen in Section 4.2.2), and adopt scalable learning techniques similar to those used for video foundation models (Section 4.1). Though a reliance on language captions may limit scalability

to an extent, this may become less of an issue as automated captioning methods improve (Section 5.1.2). Second, we should focus on methods that can best obtain the key LfV benefits (e.g., generalisation beyond the available robot data). This includes targeting the learning of policies (Section 4.3.1) and dynamics models (Section 4.3.2) from video data. Improved benchmarks that quantitatively measure LfV benefits (see Section 5.2.2) will facilitate these efforts.

- *Action representations.* In Section 4.2.1, we outlined methods for extracting action representations from video. These can serve to mitigate missing action labels in video. These methods are promising, but further work is required to scale them to large, realistic internet video data.
- *Improved datasets.* Improving and scaling our video datasets (as per the desiderata outlined in Section 5.1.1) is a reliable way to improve our LfV models. Suitable methods for curating internet video data are outlined in Section 5.1.2.
- *Is scaling enough?* Exploiting internet video will likely drive significant advances in robotics. However, the generalist robot setting presents challenges that may not be solved via naive scaling of deep learning (Section 6.4). As LfV and other research attempts to address the data bottleneck in robotics, other complementary and alternative research agendas should be explored in parallel (see Section 2).

The analyses, taxonomies, and directions presented in this survey should serve as a valuable reference for future LfV research. We hope this can catalyze further research in the area, and accelerate our progress towards developing general-purpose robots.

Acknowledgments

Daniel C.H. Tan is supported by the Agency for Science, Technology, and Research (A*STAR), Singapore. Fernando Acero and Nathan Herr are supported by the UKRI CDT in Foundational AI [EP/S021566/1] at the UCL Centre for Artificial Intelligence. We thank Anthony Hu for useful feedback in the early stages of the paper.

References

[1] D. Abel, A. Barreto, B. Van Roy, D. Precup, H. P. van Hasselt, and S. Singh. 2024. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

[2] J. Achiam et al. 2023. Gpt-4 technical report. (2023).

[3] A. Adeniji, A. Xie, C. Sferrazza, Y. Seo, S. James, and P. Abbeel. 2023. Language reward modulation for pretraining reinforcement learning. *ArXiv*, abs/2308.12270. https://api.semanticscholar.org/CorpusID:261075941.

[4] M. Ahn, H. Zhu, K. Hartikainen, H. Ponte, A. Gupta, S. Levine, and V. Kumar. 2020. Robel: robotics benchmarks for learning with low-cost robots. In *Conference on robot learning*. PMLR, 1300–1313.

[5] M. Ahn et al. 2024. Autort: embodied foundation models for large scale orchestration of robotic agents. (2024).

[6] M. Ahn et al. 2022. Do as i can, not as i say: grounding language in robotic affordances. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:247939706.

[7] A. Ajay et al. 2023. Compositional foundation models for hierarchical planning. *ArXiv*, abs/2309.08587. https://api.semanticscholar.org/CorpusID:262012485.

[8] I. Akkaya et al. 2019. Solving rubik's cube with a robot hand. (2019).

[9] J.-B. Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35, 23716–23736.

[10] S. Ali, O. Hassan, A. Gopalakrishnan, A. S. V. Muriyan, and S. L. Francis. 2022. Unmanned aerial vehicles: a literature review. *Journal of Hunan University Natural Sciences*. https://api.semanticscholar.org/CorpusID:252183358.

[11] U. Anwar et al. 2024. Foundational challenges in assuring alignment and safety of large language models. (2024).

[12] Y. Aytar, T. Pfaff, D. Budden, T. L. Paine, Z. Wang, and N. de Freitas. 2018. Playing hard exploration games by watching youtube. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:44061126.

[13] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. 2017. Stochastic variational video prediction. (2017).

[14] S. Bahl, A. Gupta, and D. Pathak. 2022. Human-to-robot imitation in the wild. *ArXiv*, abs/2207.09450. https://api.semanticscholar.org/CorpusID:248941578.

[15] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. 2023. Affordances from human videos as a versatile representation for robotics. (2023). https://api.semanticscholar.org/CorpusID:258180471.

[16] Y. Bai et al. 2022. Constitutional ai: harmlessness from ai feedback. (2022).

[17] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. 2021. Frozen in time: a joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

[18] B. Baker, I. Akkaya, P. Zhokhov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune. 2022. Video pretraining (vpt): learning to act by watching unlabeled online videos. *ArXiv*, abs/2206.11795. https://api.semanticscholar.org/CorpusID:249953673.

[19] I. Balažević, Y. Shi, P. Papalampidi, R. Chaabouni, S. Koppula, and O. J. Hénaff. 2024. Memory consolidation enables long-context video understanding. (2024).

[20] O. Bar-Tal et al. 2024. Lumiere: a space-time diffusion model for video generation. (2024).

[21] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. 2023. V-jepa: latent video prediction for visual representation learning. (2023).

[22] K. Baumli et al. 2023. Vision-language models as a source of rewards. (2023).

[23] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh. 2024. Rt-h: action hierarchies using language. (2024).

[24] J. Betker et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2, 3, 8.

[25] H. Bharadhwaj. 2024. Position: scaling simulation is neither necessary nor sufficient for in-the-wild robot manipulation. In *Forty-first International Conference on Machine Learning*.

[26] H. Bharadhwaj, A. Gupta, V. Kumar, and S. Tulsiani. 2023. Towards generalizable zero-shot manipulation via translating human interaction plans. *ArXiv*, abs/2312.00775. https://api.semanticscholar.org/CorpusID:265551754.

[27] C. Bhateja, D. Guo, D. Ghosh, A. Singh, M. Tomar, Q. H. Vuong, Y. Chebotar, S. Levine, and A. Kumar. 2023. Robotic offline rl from internet videos via value-function pre-training. *ArXiv*, abs/2309.13041. https://api.semanticscholar.org/CorpusID:262217278.

[28] P. Biswal and P. K. Mohanty. 2020. Development of quadruped walking robots: a review. *Ain Shams Engineering Journal*. https://api.semanticscholar.org/CorpusID:229445610.

[29] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. 2023. Training diffusion models with reinforcement learning. (2023).

[30] K. Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, and S. Levine. 2023. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *ArXiv*, abs/2310.10639. https://api.semanticscholar.org/CorpusID:264172455.

[31] N. Blank, M. Reuss, F. Wenzel, O. Mees, and R. Lioutikov. 2024. Scaling robot policy learning via zero-shot labeling with foundation models. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*.

[32] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. 2023. Align your latents: high-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

[33] A. Blattmann et al. 2023. Stable video diffusion: scaling latent video diffusion models to large datasets. (2023).

[34] R. Bommasani et al. 2021. On the opportunities and risks of foundation models. (2021).

[35] K. Bousmalis et al. 2023. Robocat: a self-improving foundation agent for robotic manipulation. (2023).

[36] A. Brohan et al. 2022. Rt-1: robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817. https://api.semanticscholar.org/CorpusID:254591260.

[37] A. Brohan et al. 2023. Rt-2: vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818. https://api.semanticscholar.org/CorpusID:260293142.

[38] T. Brooks et al. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators.

[39] T. B. Brown et al. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165. https://api.semanticscholar.org/CorpusID:218971783.

[40] J. Bruce et al. 2024. Genie: generative interactive environments. In https://api.semanticscholar.org/CorpusID:267897982.

[41] K. Burns, Z. Witzel, J. I. Hamid, T. Yu, C. Finn, and K. Hausman. 2023. What makes pre-trained visual representations successful for robust manipulation? *ArXiv*, abs/2312.12444. https://api.semanticscholar.org/CorpusID:266369884.

[42] S. Cai, B. Zhang, Z. Wang, X. Ma, A. Liu, and Y. Liang. 2023. Groot: learning to follow instructions by watching gameplay videos. *ArXiv*, abs/2310.08235. https://api.semanticscholar.org/CorpusID:263908999.

[43] E. Chane-Sane, C. Schmid, and I. Laptev. 2023. Learning video-conditioned policies for unseen manipulation tasks. (2023). https://api.semanticscholar.org/CorpusID:258588267.

[44] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. 2022. Maskgit: masked generative image transformer. (2022). https://api.semanticscholar.org/CorpusID:246680316.

[45] M. Chang, A. Gupta, and S. Gupta. 2022. Learning value functions from undirected state-only experience. *ArXiv*, abs/2204.12458. https://api.semanticscholar.org/CorpusID:245064979.

[46] M. Chang, A. Gupta, and S. Gupta. 2020. Semantic visual navigation by watching youtube videos. *ArXiv*, abs/2006.10034. https://api.semanticscholar.org/CorpusID:219721405.

[47] M. Chang, A. Prakash, and S. Gupta. 2023. Look ma, no hands! agent-environment factorization of egocentric videos. *ArXiv*, abs/2305.16301. https://api.semanticscholar.org/CorpusID:258888066.

[48] D. S. Chaplot, R. Salakhutdinov, A. K. Gupta, and S. Gupta. 2020. Neural topological slam for visual navigation. (2020). https://api.semanticscholar.org/CorpusID:214754592.

[49] Y. Chebotar et al. 2023. Q-transformer: scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*. PMLR, 3909–3928.

[50] A. S. Chen, S. Nair, and C. Finn. 2021. Learning generalizable robotic reward functions from "in-the-wild" human videos. *ArXiv*, abs/2103.16817. https://api.semanticscholar.org/CorpusID:232428118.

[51] J. Chen, D. Zhu, K. Haydarov, X. Li, and M. Elhoseiny. 2023. Video chatcaptioner: towards the enriched spatiotemporal descriptions. (2023).

[52] X. Chen, S. Xie, and K. He. 2021. An empirical study of training self-supervised vision transformers. (2021). https://api.semanticscholar.org/CorpusID:233024948.

[53] Z. Chen et al. 2024. Rh20t-p: a primitive-level robotic dataset towards composable generalization agents. *arXiv preprint arXiv:2403.19622*.

[54] E. Chisari, N. Heppert, M. Argus, T. Welschehold, T. Brox, and A. Valada. 2024. Learning robotic manipulation policies from point clouds with conditional flow matching. (2024).

[55] Z. J. Cui, Y. Wang, N. M. (Shafiullah, and L. Pinto. 2022. From play to policy: conditional behavior generation from uncurated robot data. *ArXiv*, abs/2210.10047. https://api.semanticscholar.org/CorpusID:252968170.

[56] X. Dai et al. 2023. Emu: enhancing image generation models using photogenic needles in a haystack. (2023).

[57] M. Dalal, A. Mandlekar, C. Garrett, A. Handa, R. Salakhutdinov, and D. Fox. 2023. Imitating task and motion planning with visuomotor transformers. (2023).

[58] D. Damen et al. 2022. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. (2022).

[59] D. Damen et al. 2018. Scaling egocentric vision: the epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, 720–736.

[60] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. Damen. 2022. Epic-kitchens visor benchmark: video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35, 13745–13758.

[61] S. Dasari, M. K. Srirama, U. Jain, and A. Gupta. 2023. An unbiased look at datasets for visuo-motor pre-training. *ArXiv*, abs/2310.09289. https://api.semanticscholar.org/CorpusID:263914653.

[62] M. Deitke et al. 2022. Procthor: large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35, 5982–5994.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[64] H. Ding, C. Liu, S. He, X. Jiang, P. H. Torr, and S. Bai. 2023. Mose: a new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20224–20234.

[65] Z. Ding, H. Luo, K. Li, J. Yue, T. Huang, and Z. Lu. 2023. Clip4mc: an rl-friendly vision-language model for minecraft. *ArXiv*, abs/2303.10571. https://api.semanticscholar.org/CorpusID:257632482.

[66] Y. Du and L. P. Kaelbling. 2024. Compositional generative modeling: a single model is not all you need. *ArXiv*, abs/2402.01103. https://api.semanticscholar.org/CorpusID:267406745.

[67] Y. Du, M. Yang, B. Dai, H. Dai, O. Nachum, J. B. Tenenbaum, D. Schuurmans, and P. Abbeel. 2023. Learning universal policies via text-guided video generation. *ArXiv*, abs/2302.00111. https://api.semanticscholar.org/CorpusID:256459809.

[68] Y. Du et al. 2023. Video language planning. *ArXiv*, abs/2310.10625. https://api.semanticscholar.org/CorpusID:264172935.

[69] Y. Du, K. Konyushkova, M. Denil, A. Raju, J. Landon, F. Hill, N. de Freitas, and S. Cabi. 2023. Vision-language models as success detectors. (2023).

[70] A. Dubey et al. 2024. The llama 3 herd of models. (2024).

[71] A. D. Edwards and C. L. Isbell. 2019. Perceptual values from observation. (2019).

[72] A. D. Edwards, H. Sahni, Y. Schroecker, and C. L. Isbell. 2018. Imitating latent policies from observation. *ArXiv*, abs/1805.07914. https://api.semanticscholar.org/CorpusID:29156793.

[73] A. D. Edwards et al. 2020. Estimating q(s, s') with deep deterministic dynamics gradients. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:211258729.

[74] A. Escontrela, A. Adeniji, W. Yan, A. Jain, X. B. Peng, K. Goldberg, Y. Lee, D. Hafner, and P. Abbeel. 2023. Video prediction models as rewards for reinforcement learning. *ArXiv*, abs/2305.14343. https://api.semanticscholar.org/CorpusID:258841355.

[75] C. Eze and C. Crick. 2024. Learning by watching: a review of video-based learning approaches for robot manipulation. (2024).

[76] M. Faldor, J. Zhang, A. Cully, and J. Clune. 2024. Omni-epic: open-endedness via models of human notions of interestingness with environments programmed in code. (2024).

[77] L. Fan et al. 2022. Minedojo: building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35, 18343–18362.
[78] Z. Gao, C. Tan, L. Wu, and S. Z. Li. 2022. Simvp: simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3180.
[79] N. Gavenski, O. Rodrigues, and M. Luck. 2024. Imitation learning: a survey of learning methods, environments and metrics. (2024).
[80] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh. 2022. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*. Springer, 102–118.
[81] S. Ge et al. 2023. Preserve your own correlation: a noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22930–22941.
[82] D. Ghosh, C. A. Bhateja, and S. Levine. 2023. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*. PMLR, 11321–11339.
[83] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. 2022. Omnimae: single model masked pretraining on images and videos. (2022). https://api.semanticscholar.org/CorpusID:249712367.
[84] I. J. Goodfellow, J. Shlens, and C. Szegedy. 2014. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572. https://api.semanticscholar.org/CorpusID:6706414.
[85] R. Goyal et al. 2017. The "something something" video database for learning and evaluating visual common sense. (2017). https://api.semanticscholar.org/CorpusID:834612.
[86] K. Grauman et al. 2023. Ego-exo4d: understanding skilled human activity from first-and third-person perspectives. (2023).
[87] K. Grauman et al. 2021. Ego4d: around the world in 3,000 hours of egocentric video. (2021). https://api.semanticscholar.org/CorpusID:238856888.
[88] A. Gu and T. Dao. 2023. Mamba: linear-time sequence modeling with selective state spaces. (2023).
[89] J. Gu et al. 2023. Maniskill2: a unified benchmark for generalizable manipulation skills. (2023).
[90] L. Guan, Y. Zhou, D. Liu, Y. Zha, H. B. Amor, and S. Kambhampati. 2024. " task success" is not enough: investigating the use of video-language models as behavior critics for catching undesirable agent behaviors. (2024).
[91] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei. 2022. Maskvit: masked visual pre-training for video prediction. (2022).
[92] H. Ha, P. Florence, and S. Song. 2023. Scaling up and distilling down: language-guided robot skill acquisition. In *Conference on Robot Learning*. PMLR, 3766–3777.
[93] T. Haarnoja et al. 2018. Soft actor-critic algorithms and applications. (2018).
[94] D. Hafner, J. Paukonis, J. Ba, and T. P. Lillicrap. 2023. Mastering diverse domains through world models. *ArXiv*, abs/2301.04104. https://api.semanticscholar.org/CorpusID:255569874.
[95] H. He, C. Bai, L. Pan, W. Zhang, B. Zhao, and X. Li. 2024. Large-scale actionless video pre-training via discrete diffusion for efficient policy learning. (2024).
[96] D. Hendrycks and T. G. Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261. https://api.semanticscholar.org/CorpusID:56657912.
[97] J. Hilton, J. Tang, and J. Schulman. 2023. Scaling laws for single-agent reinforcement learning. *ArXiv*, abs/2301.13442. https://api.semanticscholar.org/CorpusID:256416224.
[98] J. Ho and S. Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.
[99] J. Ho, A. Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239. https://api.semanticscholar.org/CorpusID:219955663.
[100] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35, 8633–8646.
[101] J. Ho et al. 2022. Imagen video: high definition video generation with diffusion models. (2022).
[102] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi. 2022. Diffusion models for video prediction and infilling. (2022).
[103] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang. 2025. Egodex: learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*.
[104] D. P. Horan, E. Richardson, and Y. Weiss. 2021. When is unsupervised disentanglement possible? In https://api.semanticscholar.org/CorpusID:248497880.
[105] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. 2023. Gaia-1: a generative world model for autonomous driving. *ArXiv*, abs/2309.17080. https://api.semanticscholar.org/CorpusID:263310665.
[106] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. Lora: low-rank adaptation of large language models. (2021).
[107] Y. Hu et al. 2023. Toward general-purpose robots via foundation models: a survey and meta-analysis. (2023).
[108] Y. Hu, R. Wang, L. Li, and Y. Gao. 2023. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:258048578.

[109] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. 2023. Large language models cannot self-correct reasoning yet. (2023).

[110] T. Huang, G. Jiang, Y. Ze, and H. Xu. 2023. Diffusion reward: learning rewards via conditional video diffusion. (2023).

[111] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. 2023. Voxposer: composable 3d value maps for robotic manipulation with language models. (2023).

[112] W. Huang et al. 2022. Inner monologue: embodied reasoning through planning with language models. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:250451569.

[113] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine. 2021. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40, 4-5, 698–721.

[114] V. Jain et al. 2024. Vid2robot: end-to-end video-conditioned policy learning with cross-attention transformers. (2024).

[115] E. Jang. 2024. All neural networks, all autonomous, all 1x speed. https://www.1x.tech/discover/all-neural-networks-all-autonomous-all-1x-speed. Accessed: 2024-04-10. (Feb. 2024).

[116] H. Jang, D. Kim, J. Kim, J. Shin, P. Abbeel, and Y. Seo. 2024. Visual representation learning with stochastic frame prediction. (2024).

[117] Y. Ji, Z. Li, Y. Sun, X. B. Peng, S. Levine, G. Berseth, and K. Sreenath. 2022. Hierarchical reinforcement learning for precise soccer shooting skills using a quadrupedal robot. (2022). https://api.semanticscholar.org/CorpusID:251253136.

[118] A. Q. Jiang et al. 2024. Mixtral of experts. (2024).

[119] Y. Jin et al. 2023. Unified language-vision pretraining with dynamic discrete visual tokenization. (2023).

[120] Y. Jing, X. Zhu, X. Liu, Q. Sima, T. Yang, Y. Feng, and T. Kong. 2023. Exploring visual pre-training for robot manipulation: datasets, models and methods. *ArXiv*, abs/2308.03620. https://api.semanticscholar.org/CorpusID:254198890.

[121] X. Ju, Y. Gao, Z. Zhang, Z. Yuan, X. Wang, A. Zeng, Y. Xiong, Q. Xu, and Y. Shan. 2024. Miradata: a large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37, 48955–48970.

[122] V. JyothirS, S. Jalagam, Y. LeCun, and V. Sobal. 2023. Gradient-based planning with world models. *ArXiv*, abs/2312.17227. https://api.semanticscholar.org/CorpusID:266573170.

[123] J. Kaplan et al. 2020. Scaling laws for neural language models. (2020).

[124] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. 2023. Cotracker: it is better to track together. (2023).

[125] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. 2023. Language-driven representation learning for robotics. *ArXiv*, abs/2302.12766. https://api.semanticscholar.org/CorpusID:257205716.

[126] H. Karnan, G. Warnell, X. Xiao, and P. Stone. 2021. Voila: visual-observation-only imitation learning for autonomous navigation. (2021). https://api.semanticscholar.org/CorpusID:234790310.

[127] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza. 2023. Champion-level drone racing using deep reinforcement learning. *Nature*, 620, 7976, 982–987.

[128] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier. 2023. A survey of reinforcement learning from human feedback. (2023).

[129] A. Khazatsky et al. 2024. Droid: a large-scale in-the-wild robot manipulation dataset. *ArXiv*, abs/2403.12945. https://api.semanticscholar.org/CorpusID:268531351.

[130] M. J. Kim, J. Wu, and C. Finn. 2023. Giving robots a hand: learning generalizable manipulation with eye-in-hand human video demonstrations. *ArXiv*, abs/2307.05959. https://api.semanticscholar.org/CorpusID:259836885.

[131] M. J. Kim et al. 2024. Openvla: an open-source vision-language-action model. (2024).

[132] A. Kirillov et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

[133] M. Klissarov, P. D'Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff. 2023. Motif: intrinsic motivation from artificial intelligence feedback. (2023).

[134] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. Tenenbaum. 2023. Learning to act from actionless videos through dense correspondences. *ArXiv*, abs/2310.08576. https://api.semanticscholar.org/CorpusID:263908842.

[135] E. Kolve et al. 2017. Ai2-thor: an interactive 3d environment for visual ai. (2017).

[136] D. Kondratyuk et al. 2023. Videopoet: a large language model for zero-shot video generation. (2023).

[137] E. Krotkov, D. Hackett, L. Jackel, M. Perschbacher, J. Pippine, J. Strauss, G. Pratt, and C. Orlowski. 2018. The darpa robotics challenge finals: results and perspectives. (2018).

[138] A. Kumar, A. Zhou, G. Tucker, and S. Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 1179–1191.

[139] N. Kumar. 2024. Will scaling solve robotics? Accessed: 2024-10-08. (2024). https://spectrum.ieee.org/solve-robotics.

[140] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang. 2022. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:251223373.

[141] V. Kumar, R. Shah, G. Zhou, V. Moens, V. Caggiano, A. Gupta, and A. Rajeswaran. 2024. Robohive: a unified framework for robot learning. *Advances in Neural Information Processing Systems*, 36.

[142] H. Küttler, N. Nardelli, A. Miller, R. Raileanu, M. Selvatici, E. Grefenstette, and T. Rocktäschel. 2020. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33, 7671–7684.

[143] L. A. Layne. 2023. Robot-related fatalities at work in the united states, 1992-2017. *American journal of industrial medicine*. https://api.semanticscholar.org/CorpusID:257370737.

[144] Y. LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 1.

[145] Y. Lee, J. J. Lim, A. Anandkumar, and Y. Zhu. 2021. Adversarial skill chaining for long-horizon robot manipulation via terminal state regularization. (2021). https://api.semanticscholar.org/CorpusID:237351177.

[146] S. Levine, A. Kumar, G. Tucker, and J. Fu. 2020. Offline reinforcement learning: tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643. https://api.semanticscholar.org/CorpusID:218486979.

[147] J. Levy, T. Westenbroek, and D. Fridovich-Keil. 2024. Learning to walk from three minutes of real-world data with semi-structured dynamics models. (2024).

[148] C. Li et al. 2024. Behavior-1k: a human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. (2024).

[149] J. Li, M. Stamatopoulou, and D. Kanoulas. 2024. Sds–see it, do it, sorted: quadruped skill synthesis from single video demonstration. (2024).

[150] J. Li et al. 2024. Numinamath tir. https://huggingface.co/AI-MO/NuminaMath-TIR. (2024).

[151] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu. 2024. Okami: teaching humanoid robots manipulation skills through single video imitation. (2024).

[152] J. Li, W. Wang, Y. Peng, C. Shen, Y. Zhu, and Z. Xu. 2024. Visual robotic manipulation with depth-aware pretraining. In *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 843–850.

[153] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao. 2023. Unmasked teacher: towards training-efficient video foundation models. (2023). https://api.semanticscholar.org/CorpusID:257771777.

[154] P. Li, T. Liu, Y. Li, M. Han, H. Geng, S. Wang, Y. Zhu, S.-C. Zhu, and S. Huang. 2024. Ag2manip: learning novel manipulation skills with agent-agnostic visual and action representations. (2024).

[155] S. Li, Y. Du, J. B. Tenenbaum, A. Torralba, and I. Mordatch. 2022. Composing ensembles of pre-trained models via iterative consensus. (2022).

[156] X. Li et al. 2024. Evaluating real-world robot manipulation policies in simulation. (2024).

[157] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. 2023. Code as policies: language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9493–9500.

[158] W. Liang, S. Wang, H.-J. Wang, O. Bastani, D. Jayaraman, and Y. J. Ma. 2024. Environment curriculum generation via large language models. (2024).

[159] S. Lifshitz, K. Paster, H. Chan, J. Ba, and S. A. McIlraith. 2023. Steve-1: a generative model for text-to-behavior in minecraft. *ArXiv*, abs/2306.00937. https://api.semanticscholar.org/CorpusID:258999563.

[160] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. 2023. Video-llava: learning united visual representation by alignment before projection. (2023).

[161] X. Lin, J. So, S. Mahalingam, F. Liu, and P. Abbeel. 2024. Spawnnet: learning generalizable visuomotor skills from pre-trained network. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4781–4787.

[162] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. 2024. Libero: benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36.

[163] H. Liu, W. Yan, M. Zaharia, and P. Abbeel. 2024. World model on million-length video and language with blockwise ringattention. (2024). https://api.semanticscholar.org/CorpusID:268385142.

[164] H. Liu, M. Zaharia, and P. Abbeel. 2023. Ring attention with blockwise transformers for near-infinite context. (2023).

[165] H. Liu, M. Zhuge, B. Li, Y. Wang, F. Faccio, B. Ghanem, and J. Schmidhuber. 2023. Learning to identify critical states for reinforcement learning from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1955–1965.

[166] J. Liu, Y. Yuan, J. Hao, F. Ni, L. Fu, Y. Chen, and Y. Zheng. 2024. Enhancing robotic manipulation with ai feedback from multimodal large language models. (2024).

[167] M. Lutter, C. Ritter, and J. Peters. 2019. Deep lagrangian networks: using physics as model prior for deep learning. (2019).

[168] C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet. 2020. Learning latent plans from play. In *Conference on robot learning*. PMLR, 1113–1132.

[169] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. 2023. Interactive language: talking to robots in real time. (2023).

[170] Y. J. Ma, W. J. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman. 2023. Liv: language-image representations and rewards for robotic control. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:258999195.

[171] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. 2022. Vip: towards universal visual reward and representation via value-implicit pre-training. *ArXiv*, abs/2210.00030. https://api.semanticscholar.org/CorpusID:252683397.

[172] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. 2023. Video-chatgpt: towards detailed video understanding via large vision and language models. (2023).

[173] N. Madan, A. Moegelmose, R. Modi, Y. S. Rawat, and T. B. Moeslund. 2024. Foundation models for video understanding: a survey. (2024).
[174] A. Majumdar et al. 2023. Where are we in the search for an artificial visual cortex for embodied intelligence? *ArXiv*, abs/2303.18240. https://api.semanticscholar.org/CorpusID:257901087.
[175] V. Makoviychuk et al. 2021. Isaac gym: high performance gpu-based physics simulation for robot learning. (2021).
[176] P. Mandikal and K. Grauman. 2022. Dexvip: learning dexterous grasping with human hand pose priors from video. *ArXiv*, abs/2202.00164. https://api.semanticscholar.org/CorpusID:237369373.
[177] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox. 2023. Mimicgen: a data generation system for scalable robot learning using human demonstrations. (2023).
[178] P. Marza, L. Matignon, O. Simonin, and C. Wolf. 2022. Multi-object navigation with dynamically learned neural implicit representations. (2022). https://api.semanticscholar.org/CorpusID:252815543.
[179] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard. 2022. Calvin: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7, 3, 7327–7334.
[180] R. Mendonca, S. Bahl, and D. Pathak. 2023. Structured world models from human videos. *ArXiv*, abs/2308.10901. https://api.semanticscholar.org/CorpusID:259336798.
[181] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. 2019. Howto100m: learning a text-video embedding by watching hundred million narrated video clips. (2019). https://api.semanticscholar.org/CorpusID:182952863.
[182] U. A. Mishra, S. Xue, Y. Chen, and D. Xu. 2023. Generative skill chaining: long-horizon skill planning with diffusion models. PMLR, (2023).
[183] G. E. Moore. 1998. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86, 1, 82–85.
[184] H. Moravec. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, Cambridge, MA.
[185] Y. Mu et al. 2023. Embodiedgpt: vision-language pre-training via embodied chain of thought. *ArXiv*, abs/2305.15021. https://api.semanticscholar.org/CorpusID:258865718.
[186] T. Nagarajan and K. Grauman. 2021. Shaping embodied agent behavior with activity-context priors from egocentric video. *ArXiv*, abs/2110.07692. https://api.semanticscholar.org/CorpusID:239009498.
[187] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*. Springer, 407–426.
[188] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. 2022. R3m: a universal visual representation for robot manipulation. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:247618840.
[189] S. Narang and A. Chowdhery. 2022. Pathways language model (palm): scaling to 540 billion parameters for breakthrough performance. (2022).
[190] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. 2024. Robocasa: large-scale simulation of everyday tasks for generalist robots. (2024).
[191] S. Nasiriany et al. 2024. Pivot: iterative visual prompting elicits actionable knowledge for vlms. *ArXiv*, abs/2402.07872. https://api.semanticscholar.org/CorpusID:267627797.
[192] A. v. d. Oord, Y. Li, and O. Vinyals. 2018. Representation learning with contrastive predictive coding. (2018).
[193] OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774. https://api.semanticscholar.org/CorpusID:257532815.
[194] L. Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35, 27730–27744.
[195] A. Padalkar et al. 2023. Open x-embodiment: robotic learning datasets and rt-x models. (2023).
[196] P. Papalampidi et al. 2023. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *ArXiv*, abs/2312.07395. https://api.semanticscholar.org/CorpusID:266174654.
[197] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. K. Gupta. 2022. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:247292805.
[198] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. 2023. Generative agents: interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.
[199] S. Park, D. Ghosh, B. Eysenbach, and S. Levine. 2024. Hiql: offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36.
[200] J. Parker-Holder, M. Jiang, M. Dennis, M. Samvelyan, J. N. Foerster, E. Grefenstette, and T. Rocktaschel. 2022. Evolving curricula with regret-based environment design. *ArXiv*, abs/2203.01302. https://api.semanticscholar.org/CorpusID:247218125.
[201] C. Peng, C. Xu, Y. Wang, M. Ding, H. Yang, M. Tomizuka, K. Keutzer, M. Pavone, and W. Zhan. 2024. Q-slam: quadric representations for monocular slam. (2024).
[202] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine. 2020. Learning agile robotic locomotion skills by imitating animals. (2020).

[203] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. 2018. Sfv: reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37, 178. https://api.semanticscholar.org/CorpusID:52937281.

[204] K. Pertsch, R. Desai, V. Kumar, F. Meier, J. J. Lim, D. Batra, and A. Rai. 2022. Cross-domain transfer via semantic skill imitation. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:254636470.

[205] K. Pertsch, O. Rybkin, J. Yang, S. Zhou, K. G. Derpanis, K. Daniilidis, J. J. Lim, and A. Jaegle. 2019. Keyframing the future: keyframe discovery for visual prediction and planning. In *Conference on Learning for Dynamics & Control*. https://api.semanticscholar.org/CorpusID:218571383.

[206] J. Peters, D. D. Lee, J. Kober, D. Nguyen-Tuong, J. A. Bagnell, and S. Schaal. 2016. Robot learning. (2016).

[207] R. F. Prudencio, M. R. Maximo, and E. L. Colombini. 2023. A survey on offline reinforcement learning: taxonomy, review, and open problems. (2023).

[208] X. Puig et al. 2023. Habitat 3.0: a co-habitat for humans, avatars and robots. (2023).

[209] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox. 2024. The colosseum: a benchmark for evaluating generalization for robotic manipulation. (2024).

[210] Y. Qin, H. Su, and X. Wang. 2022. From one hand to multiple hands: imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7, 10873–10881. https://api.semanticscholar.org/CorpusID:248392006.

[211] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. 2021. Dexmv: imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:236986915.

[212] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1, 8, 9.

[213] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. 2022. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:252718704.

[214] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik. 2024. Humanoid locomotion as next token prediction. (2024).

[215] K. Ranasinghe, X. Li, K. Kahatapitiya, and M. S. Ryoo. 2024. Understanding long videos in one multimodal language model pass. (2024).

[216] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard. 2020. Recent advances in robot learning from demonstration. *Annu. Rev. Control. Robotics Auton. Syst.*, 3, 297–330. https://api.semanticscholar.org/CorpusID:208958394.

[217] S. Reed et al. 2022. A generalist agent. *Trans. Mach. Learn. Res.*, 2022. https://api.semanticscholar.org/CorpusID:248722148.

[218] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2021. High-resolution image synthesis with latent diffusion models. (2021). https://api.semanticscholar.org/CorpusID:245335280.

[219] Y. Rong, T. Shiratori, and H. Joo. 2020. Frankmocap: fast monocular 3d hand and body motion capture by regression and integration. (2020).

[220] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard. 2022. Latent plans for task-agnostic offline reinforcement learning. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:252367910.

[221] O. Rybkin, K. Pertsch, K. G. Derpanis, K. Daniilidis, and A. Jaegle. 2018. Learning what you can do before doing anything. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:60441438.

[222] S. Sartor and N. Thompson. 2024. Neural scaling laws for embodied ai. *ArXiv*, abs/2405.14005. https://api.semanticscholar.org/CorpusID:269983385.

[223] M. C. Schiappa, Y. S. Rawat, and M. Shah. 2023. Self-supervised learning for videos: a survey. *ACM Computing Surveys*, 55, 13s, 1–37.

[224] K. Schilling and C. C. Jungius. 1995. Mobile robots for planetary exploration. *Control Engineering Practice*, 4, 513–524. https://api.semanticscholar.org/CorpusID:109713134.

[225] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn. 2020. Reinforcement learning with videos: combining offline observations with interaction. *ArXiv*, abs/2011.06507. https://api.semanticscholar.org/CorpusID:226306712.

[226] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Daniilidis, S. Levine, and C. Finn. 2019. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:209515451.

[227] D. Schmidt and M. Jiang. 2023. Learning to act without actions. *ArXiv*, abs/2312.10812. https://api.semanticscholar.org/CorpusID:266359570.

[228] J. L. Schonberger and J.-M. Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

[229] J. Schrittwieser et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588, 7839, 604–609.

[230] C. Schuhmann et al. 2022. Laion-5b: an open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35, 25278–25294.

[231] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal policy optimization algorithms. (2017).

[232] Y. Seo, K. Lee, S. James, and P. Abbeel. 2022. Reinforcement learning with action-free pre-training from videos. *ArXiv*, abs/2203.13880. https://api.semanticscholar.org/CorpusID:247762941.

[233] Y. Seo, K. Lee, F. Liu, S. James, and P. Abbeel. 2022. Harp: autoregressive latent video prediction with high-fidelity image generator. (2022). https://api.semanticscholar.org/CorpusID:252280733.

[234] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. 2018. Time-contrastive networks: self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 1134–1141.

[235] P. Sermanet et al. 2023. Robovqa: multimodal long-horizon reasoning for robotics. (2023).

[236] C. Sferrazza, D.-M. Huang, F. Liu, J. Lee, and P. Abbeel. 2024. Body transformer: leveraging robot embodiment for policy learning. (2024).

[237] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto. 2023. On bringing robots home. (2023).

[238] D. Shah, B. Osiński, S. Levine, et al. 2023. Lm-nav: robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*. PMLR, 492–504.

[239] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. 2020. Understanding human hands in contact at internet scale. (2020). https://api.semanticscholar.org/CorpusID:215413188.

[240] C. Shang, A. You, S. Subramanian, T. Darrell, and R. Herzig. 2024. Traveler: a multi-lmm agent framework for video question-answering. (2024).

[241] J. Shang and M. S. Ryoo. 2021. Self-supervised disentangled representation learning for third-person imitation learning. (2021). https://api.semanticscholar.org/CorpusID:236772575.

[242] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. 2020. Concept2robot: learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40, 1419–1434. https://api.semanticscholar.org/CorpusID:220069237.

[243] K. Shaw, S. Bahl, and D. Pathak. 2022. Videodex: learning dexterity from internet videos. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:254408735.

[244] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn. 2024. Yell at your robot: improving on-the-fly from language corrections. (2024).

[245] M. Sieb, Z. Xian, A. Huang, O. Kroemer, and K. Fragkiadaki. 2020. Graph-structured visual imitation. In *Conference on Robot Learning*. PMLR, 979–989.

[246] S. Silwal et al. 2023. What do we learn from a large-scale study of pre-trained visual representations in sim and real environments? *ArXiv*, abs/2310.02219. https://api.semanticscholar.org/CorpusID:263608779.

[247] U. Singer et al. 2022. Make-a-video: text-to-video generation without text-video data. (2022).

[248] A. Sivakumar, K. Shaw, and D. Pathak. 2022. Robotic telekinesis: learning a robotic hand imitator by watching humans on youtube. *ArXiv*, abs/2202.10448. https://api.semanticscholar.org/CorpusID:247011104.

[249] A. Sjöberg. 2023. How many videos are there on youtube? https://www.dexerto.com/entertainment/how-many-videos-are-there-on-youtube-2197264/. (Dec. 2023).

[250] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. 2019. Avid: learning multi-stage tasks via pixel-level translation of human videos. *ArXiv*, abs/1912.04443. https://api.semanticscholar.org/CorpusID:209140723.

[251] A. Sohn et al. 2024. Introducing rfm-1: giving robots human-like reasoning capabilities. https://covariant.ai/insights/introducing-rfm-1-giving-robots-human-like-reasoning-capabilities/. Accessed: 2024-03-29. (Mar. 2024).

[252] W. Song, H. Zhao, P. Ding, C. Cui, S. Lyu, Y. Fan, and D. Wang. 2024. Germ: a generalist robotic model with mixture-of-experts for quadruped robot. (2024).

[253] S. A. Sontakke, J. Zhang, S. M. R. Arnold, K. Pertsch, E. Biyik, D. Sadigh, C. Finn, and L. Itti. 2023. Roboclip: one demonstration is enough to learn robot policies. *ArXiv*, abs/2310.07899. https://api.semanticscholar.org/CorpusID:263909538.

[254] B. C. Stadie, P. Abbeel, and I. Sutskever. 2017. Third-person imitation learning. (2017).

[255] J. C. Stroud, Z. Lu, C. Sun, J. Deng, R. Sukthankar, C. Schmid, and D. A. Ross. 2020. Learning video representations from textual web supervision. (2020).

[256] R. Sutton. 2019. The bitter lesson. (2019). http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[257] R. S. Sutton and A. G. Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[258] A. C. Tam, N. C. Rabinowitz, A. K. Lampinen, N. A. Roy, S. C. Y. Chan, D. Strouse, J. X. Wang, A. Banino, and F. Hill. 2022. Semantic exploration from language abstractions and pretrained representations. *ArXiv*, abs/2204.05080. https://api.semanticscholar.org/CorpusID:248085427.

[259] B. Tang et al. 2024. Automate: specialist and generalist assembly policies over diverse geometries. (2024).

[260] C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone. 2024. Deep reinforcement learning for robotics: a survey of real-world successes. (2024).

[261] Y. Tang et al. 2023. Video understanding with large language models: a survey. (2023).

[262] A. A. Team et al. 2023. Human-timescale adaptation in an open-ended task space. (2023).

[263] G. Team et al. 2023. Gemini: a family of highly capable multimodal models. (2023).

[264] O. M. Team et al. 2023. Octo: an open-source generalist robot policy. (2023).

[265] G. Thomas, C.-A. Cheng, R. Loynd, V. Vineet, M. Jalobeanu, and A. Kolobov. 2023. Plex: making the most of the available data for robotic manipulation pretraining. *ArXiv*, abs/2303.08789. https://api.semanticscholar.org/CorpusID:257532588.

[266] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. (2017). https://api.semanticscholar.org/CorpusID:2413610.

[267] E. Todorov, T. Erez, and Y. Tassa. 2012. Mujoco: a physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 5026–5033.

[268] M. Tomar, D. Ghosh, V. Myers, A. Dragan, M. E. Taylor, P. Bachman, and S. Levine. 2023. Video-guided skill discovery. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*.

[269] Z. Tong, Y. Song, J. Wang, and L. Wang. 2022. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602. https://api.semanticscholar.org/CorpusID:247619234.

[270] F. Torabi, G. Warnell, and P. Stone. 2018. Behavioral cloning from observation. In *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:23206414.

[271] F. Torabi, G. Warnell, and P. Stone. 2018. Generative adversarial imitation from observation. *ArXiv*, abs/1807.06158. https://api.semanticscholar.org/CorpusID:49863329.

[272] F. Torabi, G. Warnell, and P. Stone. 2019. Recent advances in imitation learning from observation. In *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:173188327.

[273] C. Urrea and J. Kern. 2025. Recent advances and challenges in industrial robotics: a systematic review of technological trends and emerging applications. *Processes*, 13, 3, 832. DOI: 10.3390/pr13030832.

[274] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[275] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan. 2022. Phenaki: variable length video generation from open domain textual description. *ArXiv*, abs/2210.02399. https://api.semanticscholar.org/CorpusID:252715594.

[276] W. Wan, H. Geng, Y. Liu, Z. Shan, Y. Yang, L. Yi, and H. Wang. 2023. Unidexgrasp++: improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3891–3902.

[277] C. Wang, X. Luo, K. W. Ross, and D. Li. 2022. Vrl3: a data-driven framework for visual deep reinforcement learning. *ArXiv*, abs/2202.10324. https://api.semanticscholar.org/CorpusID:247011862.

[278] C. Wang, L. (Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. 2023. Mimicplay: long-horizon imitation learning by watching human play. *ArXiv*, abs/2302.12422. https://api.semanticscholar.org/CorpusID:257205825.

[279] J. Wang, S. Dasari, M. K. Srirama, S. Tulsiani, and A. Gupta. 2023. Manipulate by seeing: creating manipulation controllers from pre-trained representations. *ArXiv*, abs/2303.08135. https://api.semanticscholar.org/CorpusID:257505038.

[280] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. 2023. Videomae v2: scaling video masked autoencoders with dual masking. (2023). https://api.semanticscholar.org/CorpusID:257805127.

[281] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, L. Yuan, and Y.-G. Jiang. 2022. Masked video distillation: rethinking masked feature modeling for self-supervised video representation learning. (2022). https://api.semanticscholar.org/CorpusID:254408955.

[282] T. Wang, R. Liao, J. Ba, and S. Fidler. 2018. Nervenet: learning structured policy with graph neural networks. (2018). https://api.semanticscholar.org/CorpusID:65051725.

[283] T. Wang, A. Torralba, P. Isola, and A. Zhang. 2023. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning*. PMLR, 36411–36430.

[284] Y. Wang et al. 2023. Internvid: a large-scale video-text dataset for multimodal understanding and generation. *ArXiv*, abs/2307.06942. https://api.semanticscholar.org/CorpusID:259847783.

[285] Y. Wang et al. 2024. Internvideo2: scaling video foundation models for multimodal video understanding. (2024).

[286] Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan. 2023. Robogen: towards unleashing infinite data for automated robot learning via generative simulation. (2023).

[287] Z. Wang, Y. Dong, O. Delalleau, J. Zeng, G. Shen, D. Egert, J. J. Zhang, M. N. Sreedhar, and O. Kuchaiev. 2024. Helpsteer2: open-source dataset for training top-performing reward models. (2024).

[288] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. 2023. Any-point trajectory modeling for policy learning. *ArXiv*, abs/2401.00025. https://api.semanticscholar.org/CorpusID:266693687.

[289] H. Wu, Y. Jing, C.-H. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong. 2023. Unleashing large-scale video generative pre-training for visual robot manipulation. *ArXiv*, abs/2312.13139. https://api.semanticscholar.org/CorpusID:266374724.

[290] J. Wu, H. Ma, C. Deng, and M. Long. 2023. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. *ArXiv*, abs/2305.18499. https://api.semanticscholar.org/CorpusID:258967679.

[291] M. Wulfmeier, A. Byravan, S. Bechtle, K. Hausman, and N. Heess. 2023. Foundations for transfer in reinforcement learning: a taxonomy of knowledge modalities. (2023).

[292] Z. Xian, T. Gervet, Z. Xu, Y.-L. Qiao, T.-H. Wang, and Y. Wang. 2023. Towards generalist robots: a promising paradigm via generative simulation. In https://api.semanticscholar.org/CorpusID:259202431.

[293] J. Xiang et al. 2024. Pandora: towards general world model with natural language actions and video states. (2024).

[294] A. Xie, L. Lee, T. Xiao, and C. Finn. 2023. Decomposing the generalization gap in imitation learning for visual robotic manipulation. (2023).

[295] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg. 2021. Learning by watching: physical imitation of manipulation skills from human videos. (2021). https://api.semanticscholar.org/CorpusID:231632575.

[296] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao. 2022. Gmflow: learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8121–8130.

[297] H. Xu, G. Ghosh, P.-Y. (Huang, D. Okhonko, A. Aghajanyan, and F. M. L. Z. C. Feichtenhofer. 2021. Videoclip: contrastive pre-training for zero-shot video-text understanding. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:238215257.

[298] M. Xu, Z. Xu, C. Chi, M. M. Veloso, and S. Song. 2023. Xskill: cross embodiment skill discovery. *ArXiv*, abs/2307.09955. https://api.semanticscholar.org/CorpusID:259982636.

[299] Z. Xu et al. 2024. Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots. (2024).

[300] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5036–5045.

[301] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu. 2022. Videococa: video-text modeling with zero-shot transfer from contrastive captioners. In https://api.semanticscholar.org/CorpusID:254535696.

[302] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. 2021. Videogpt: video generation using vq-vae and transformers. *ArXiv*, abs/2104.10157. https://api.semanticscholar.org/CorpusID:233307257.

[303] C. Yang, K. Yuan, Q. Zhu, W. Yu, and Z. Li. 2020. Multi-expert learning of adaptive legged locomotion. *Science Robotics*, 5, 49, eabb2174.

[304] J. Yang, M. S. Mark, B. Vu, A. Sharma, J. Bohg, and C. Finn. 2023. Robot fine-tuning made easy: pre-training rewards and policies for autonomous real-world reinforcement learning. (2023).

[305] M. Yang, Y. Du, B. Dai, D. Schuurmans, J. B. Tenenbaum, and P. Abbeel. 2023. Probabilistic adaptation of text-to-video models. *ArXiv*, abs/2306.01872. https://api.semanticscholar.org/CorpusID:259075709.

[306] M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel. 2023. Learning interactive real-world simulators. *ArXiv*, abs/2310.06114. https://api.semanticscholar.org/CorpusID:263830899.

[307] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans. 2023. Foundation models for decision making: problems, methods, and opportunities. *ArXiv*, abs/2303.04129. https://api.semanticscholar.org/CorpusID:257378587.

[308] S. Yang, J. Walker, J. Parker-Holder, Y. Du, J. Bruce, A. Barreto, P. Abbeel, and D. Schuurmans. 2024. Video as the new language for real-world decision making. (2024).

[309] Z. Yang, K. Chen, Y. Long, and Q. Dou. 2024. Simendogs: efficient data-driven scene simulation using robotic surgery videos via physics-embedded 3d gaussians. (2024). https://api.semanticscholar.org/CorpusID:269502486.

[310] S. Ye et al. 2024. Latent action pretraining from videos. (2024). https://api.semanticscholar.org/CorpusID:273351190.

[311] W. Ye, Y. Zhang, M. Wang, S. Wang, X. Gu, P. Abbeel, and Y. Gao. 2023. Foundation reinforcement learning: towards embodied generalist agents with foundation prior assistance. *ArXiv*, abs/2310.02635. https://api.semanticscholar.org/CorpusID:263620344.

[312] L. Yu et al. 2023. Language model beats diffusion – tokenizer is key to visual generation. In https://api.semanticscholar.org/CorpusID:263830733.

[313] L. Yu et al. 2022. Magvit: masked generative video transformer. (2022). https://api.semanticscholar.org/CorpusID:254563906.

[314] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. 2020. Meta-world: a benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*. PMLR, 1094–1100.

[315] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, C. Finn, and T. Ma. 2020. Mopo: model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33, 14129–14142.

[316] T. Yu et al. 2023. Scaling robot learning with semantically imagined experience. (2023).

[317] W. Yu et al. 2023. Language to rewards for robotic skill synthesis. (2023).

[318] C. Yuan, C. Wen, T. Zhang, and Y. Gao. 2024. General flow as foundation affordance for scalable robot learning. *ArXiv*, abs/2401.11439. https://api.semanticscholar.org/CorpusID:267069070.

[319] H. Yuan, R. Wu, A. Zhao, H. Zhang, Z. Ding, and H. Dong. 2021. Dmotion: robotic visuomotor control with unsupervised forward model learned from videos. (2021). https://api.semanticscholar.org/CorpusID:232146710.

[320] K. Yuan, N. Sajid, K. Friston, and Z. Li. 2023. Hierarchical generative modelling for autonomous robots. *Nature Machine Intelligence*, 5, 12, 1402–1414.

[321] K. Zakka, A. Zeng, P. R. Florence, J. Tompson, J. Bohg, and D. Dwibedi. 2021. Xirl: cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*. https://api.semanticscholar.org/CorpusID:235368061.

[322] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. 2021. Merlot: multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 23634–23651.
[323] A. Zeng et al. 2022. Socratic models: composing zero-shot multimodal reasoning with language. (2022).
[324] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. 2024. Mm-llms: recent advances in multimodal large language models. (2024).
[325] H. Zhang, X. Li, and L. Bing. 2023. Video-llama: an instruction-tuned audio-visual language model for video understanding. (2023).
[326] L. Zhang, S. Zhou, S. Stent, and J. Shi. 2022. Fine-grained egocentric hand-object segmentation: dataset, model, and applications. In *European Conference on Computer Vision*. Springer, 127–145.
[327] L. Zhao et al. 2024. Videoprism: a foundational visual encoder for video understanding. In https://api.semanticscholar.org/CorpusID:267760035.
[328] T. Zhao, S. Karamcheti, T. Kollar, C. Finn, and P. Liang. 2022. What makes representation learning from videos hard for control? In https://api.semanticscholar.org/CorpusID:252635608.
[329] W. Zhao, J. P. Queralta, and T. Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 737–744.
[330] Z. Zhao, W. S. Lee, and D. Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 31967–31987.
[331] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 2024. 3d-vla: a 3d vision-language-action generative world model. (2024).
[332] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng. 2022. Magicvideo: efficient video generation with latent diffusion models. (2022).
[333] S. Zhou, Y. Du, J. Chen, Y. Li, D.-Y. Yeung, and C. Gan. 2024. Robodreamer: learning compositional world models for robot imagination. (2024).
[334] W. Zhou, S. Bajracharya, and D. Held. 2021. Plas: latent action space for offline reinforcement learning. In *Conference on Robot Learning*. PMLR, 1719–1735.
[335] D. Zhu, Y. Wang, J. Schmidhuber, and M. Elhoseiny. 2023. Guiding online reinforcement learning with action-free offline pretraining. *ArXiv*, abs/2301.12876. https://api.semanticscholar.org/CorpusID:256390557.
[336] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao. 2023. Robot parkour learning. (2023).

Received 6 June 2024; revised 8 November 2024; accepted 25 March 2025