From Human Videos to Robot Manipulation: A Survey on Scalable Vision-Language-Action Learning with Human-Centric Data

Zhiyuan Feng[1,5*], Qixiu Li[1,5*], Huizhi Liang[1,5*], Rushuai Yang[2,5*], Yichao Shen[3,5*], Zhiying Du[4,5*], Zhaowei Zhang[5,6*], Yu Deng[5†], Li Zhao[5], Hao Zhao[1], Zongqing Lu[6], Oier Mees[7], Marc Pollefeys[7], Jiaolong Yang[5†], Baining Guo[5]

[1]Tsinghua University
[2]HKUST
[3]Xi'an Jiaotong University
[4]Fudan University
[5]Microsoft Research Asia
[6]Peking University
[7]Microsoft Zurich

Abstract

Recent progress in generalizable embodied control has been driven by large-scale pretraining of Vision–Language–Action (VLA) models. However, most existing approaches rely on large collections of robot demonstrations, which are costly to obtain and tightly coupled to specific embodiments. Human videos, by contrast, are abundant and capture rich interactions, providing diverse semantic and physical cues for real-world manipulation. Yet, embodiment differences and the frequent absence of task-aligned annotations make their direct use in VLA models challenging. This survey provides a unified view of how human videos are transformed into effective knowledge for VLA models. We categorize existing approaches into four classes based on the action-related information they derive: (i) latent action representations that encode inter-frame changes; (ii) predictive world models that forecast future frames; (iii) explicit 2D supervision that extracts image-plane cues; and (iv) explicit 3D reconstruction that recovers geometry or motion. Beyond this taxonomy, we highlight three key open challenges in this area: structuring unstructured videos into training-ready episodes, grounding video-derived supervision into robot-executable actions under embodiment and viewpoint heterogeneity, and designing evaluation protocols that better predict real-world deployment performance and transfer efficiency, thereby informing future research directions.

Figure 1: **Overview of scalable representation-bridging paradigms for VLA models.** To leverage internet-scale human video data (top), existing methods bridge the embodiment gap via four key categories: **Latent Action**, **World Model**, **Explicit 2D**, and **Explicit 3D**. These representations transform diverse human videos into action-relevant learning signals, enabling VLA models to generate executable robot actions from image and instruction inputs (bottom).

1 Introduction

Embodied robotic tasks require agents to perceive the world, interpret instructions, and produce temporally coherent actions under physical constraints. Recent progress in VLA models suggests that large-scale multimodal pretraining can yield increasingly generalizable robotic policies, echoing the "foundation model" paradigm in language and vision [Brohan *et al.*, 2023; Kim *et al.*, 2024; Black *et al.*, 2024;

*Work done during internship at Microsoft Research Asia
†Corresponding authors

Bjorck *et al.*, 2025]. However, scaling VLA models is bottlenecked by the availability of grounded robot control data: robot demonstrations provide grounded control signals but are costly, safety-constrained, and strongly tied to a specific embodiment, sensor suite, and control frequency. Even with recent efforts to scale robot data [O'Neill *et al.*, 2024; Khazatsky *et al.*, 2024; AgiBot-World-Contributors, 2025], robot-only datasets still under-represent long-horizon task structure, rare failure modes, and the distribution shifts in objects and environments encountered at deployment, which limits the robustness and generalization of VLA policies.

To address this bottleneck, human videos offer an appealing alternative. Compared to robot demonstrations, human videos are available at vastly larger scale—ranging from massive web collections [Miech *et al.*, 2019] to diverse activities captured via lightweight wearable devices [Grauman *et al.*, 2022; Grauman *et al.*, 2024]. Beyond quantity, these resources provide rich semantic and physical cues relevant to embodied learning, such as egocentric object manipulations [Liu *et al.*, 2022; Damen *et al.*, 2020], multi-view procedural assemblies [Sener *et al.*, 2022], and controlled interactions with commonsense physics [Goyal *et al.*, 2017]. However, human videos are not "action data" for robots, they lack aligned, robot-executable action labels and proprioceptive context, and human motion does not directly map to robot kinematics and control interface. This challenge motivates the central inquiry of this survey: *when human videos enter a VLA training pipeline, what type of information do they become, and how does that information interface with policy learning?*

To bridge the gap between passive human observation and active robot control, the core challenge lies in translating the rich interaction knowledge embedded in human videos into action-relevant representations compatible with robotic policy learning. In this survey, we systematize this rapidly evolving field by categorizing methods based on the **nature of the constructed representations** along the processing pipeline. From this perspective, we identify four distinct routes: (i) **Latent Action Abstraction**, which circumvents explicit alignment by encoding inter-frame changes or intents into discrete codes or continuous embeddings [Ye *et al.*, 2025; Chen *et al.*, 2024]; (ii) **Predictive World Modeling**, which uses video as a source of predictive learning signal to forecast future outcomes and distill these representations into policies [Wu *et al.*, 2024; Bharadhwaj *et al.*, 2025]; (iii) **Explicit 2D Cues**, which extract interpretable image-plane signals—such as point tracks, masks, or flow—using off-the-shelf vision tools [Wen *et al.*, 2024; Yang *et al.*, 2025a; Tan *et al.*, 2025]; and (iv) **Explicit 3D Structure**, which recovers 3D structure (poses/trajectories) to retarget motion into robot-compatible action spaces [Yang *et al.*, 2025b; Luo *et al.*, 2025]. This taxonomy highlights the diverse *representation bridges* researchers build to convert passive video observations into active control signals.

Building on this taxonomy, our survey makes three contributions:

1. **Signal-centric taxonomy.** We introduce a pipeline-based taxonomy of *representation bridges* that transform human videos into action-relevant representations for VLA learning, and compare the four routes—latent actions, world modeling, explicit 2D cues, and explicit 3D structure—in terms of their representation form, scalability, and grounding requirements.
2. **Dataset map and available signals.** We systematize representative human-video datasets along two axes—the availability of explicit geometric/3D signals and scripted vs. in-the-wild collection—and summarize how different supervision affordances (RGB-only vs. 2D/3D signals) support different transfer pipelines.
3. **Challenges at three interfaces.** We distill open problems at three critical interfaces in the human-video-to-robot pipeline: scalable episodization of unstructured videos, heterogeneity-aware grounding under embodiment/viewpoint mismatch, and deployment-predictive evaluation for transfer efficiency.

2 Background

VLA learning addresses the challenge of mapping multimodal observations to executable actions for embodied agents. At each time step t, the agent receives a current visual observation o_t, a proprioceptive state s_t (e.g., joint states and end-effector status), and a language instruction l. Modern VLA policies typically adopt a sequence-modeling formulation to predict a chunk of future actions $a_{t:t+H}$ over a horizon H, promoting temporal consistency. This process is commonly formulated as maximum-likelihood imitation learning over a dataset of robot trajectories $\mathcal{D}_{\text{robot}}$, where each trajectory τ contains aligned sequences $(o_t, s_t, l, a_t)_{t=1}^{T}$.:

$$\max_{\theta} \sum_{\tau \in \mathcal{D}_{\text{robot}}} \sum_{t} \log \pi_\theta(a_{t:t+H} \mid o_t, s_t, l). \tag{1}$$

In practice, the action a_t is instantiated as physically grounded control targets, e.g., end-effector motion in $\mathrm{SE}(3)$ (6-DoF pose or twist), gripper open/close (binary or continuous), arm joint commands, and, for dexterous hands, high-dimensional hand pose parameters (finger joint angles) that specify 3D interaction-relevant motion. Standard approaches derive such supervision from large-scale robot datasets like Open X-Embodiment [O'Neill *et al.*, 2024], DROID [Khazatsky *et al.*, 2024], and AgiBot World [AgiBot-World-Contributors, 2025]. However, collecting robot demonstrations at scale is prohibitively expensive and often constrained by specific hardware setups, creating a bottleneck for generalizable policy learning.

To overcome this data scarcity, the field has increasingly looked toward human videos as an abundant source of interaction observations. Before large-scale end-to-end VLAs became prevalent, foundational works explored leveraging human videos primarily for auxiliary objectives or specialized domains. These approaches utilized human datasets to learn robust visual representations [Nair *et al.*, 2022; Xiao *et al.*, 2022], define reward functions for reinforcement learning [Ma *et al.*, 2023], or extract interaction affordances to guide planning [Bahl *et al.*, 2023; Borja-Diaz *et al.*, 2022; Mees *et al.*, 2023]. In the specific context of dexterous manipulation, methods like DexMV [Qin *et al.*, 2022] and

DexVIP [Mandikal and Grauman, 2022] bridged the gap by extracting 3D hand meshes for motion retargeting or utilizing hand pose priors to guide grasping policies. While these methods successfully demonstrated the value of human data, they typically rely on modular pipelines, specific morphologies, or decouple representation learning from control. Moving beyond these auxiliary or task-specific uses, this survey focuses on the emerging paradigm of *human-video-based VLA learning*, which constructs surrogate supervision from unlabelled videos and motivates a *signal-centric* taxonomy of representation bridges.

3 From Video to Action: A Taxonomy of Representation Bridges

Effectively learning VLA models from human videos requires transforming video information into *action representations*, ranging from 3D hand pose to latent vectors or internal features, that distill control-relevant dynamics for optimizing Eq. (1). Under this view, we categorize prior methods into four paradigms by representation form (Fig. 2, Table 1).

3.1 Latent Actions

Latent-action learning methods aim to derive compact action representations from videos in a self-supervised manner, thereby circumventing the high cost of acquiring additional action annotations. A common paradigm is to take observations at two time steps, o_t and o_{t+H} as input and infer a latent vector z_t, such that the future observation o_{t+H} can be reconstructed from z_t and o_t. By introducing a strong information bottleneck (*e.g.*, a VQ-VAE–style discretization [Van Den Oord *et al.*, 2017]), z_t is forced to capture only the most salient inter-frame changes associated with action-relevant dynamics, serving as a proxy for the action variable in Eq. (1) for absorbing action knowledge from human videos. Since its introduction in VLA frameworks such as **LAPA** [Ye *et al.*, 2025], key challenges have been: (i) learning a compact, expressive representation of action-relevant dynamics that generalizes across diverse embodiments and data sources; and (ii) effectively integrating latent action prediction into the VLA training pipeline.

For learning latent actions, **IGOR** [Chen *et al.*, 2024] applies asymmetric cropping augmentations during encoding and decoding to discourage incorporating absolute positional information. **UniVLA** [Bu *et al.*, 2025] replaces pixel-space supervision with DINOv2 [Oquab *et al.*, 2023] features to avoid learning noisy details, and introduces a two-stage discretization with the aid of task instruction to disentangle task-irrelevant motions (*e.g.*, camera motions). **ViPRA** [Routray *et al.*, 2025] enforces optical flow consistency loss during reconstruction to encourage physically plausible motion encoding, whereas **Motus** [Bi *et al.*, 2025a] takes optical flow between observations as direct input, explicitly reducing the influence of appearance variations in latent action encoding. Some methods further incorporate robot data with ground-truth actions to better align latent actions with robot control. For example, **villa-X** [Chen *et al.*, 2025b] jointly learns human and robot latent action spaces conditioned on embodiment context, leveraging labeled robot data to ground the latent space in action-conditioned dynamics. Similarly, **CLAP** [Zhang *et al.*, 2026] aligns human-video latents with a quantized robot-derived latent space via contrastive learning, mitigating entanglement with irrelevant visual changes. Moving beyond discrete tokens, **LAWM** [Garrido *et al.*, 2026] adopts continuous actions with sparsity constraints to better capture large-scale, in-the-wild action dynamics.

From another perspective, several works explore how to integrate latent action prediction into VLA training. **Moto** [Chen *et al.*, 2025c] adopts an autoregressive pretraining scheme for next latent action prediction, followed by inserting additional query tokens for robot action decoding during finetuning. **GO-1** [AgiBot-World-Contributors, 2025] employs a hierarchical design, learning a latent planner that predicts future latent actions to bridge the vision-language backbone and a flow-based action expert. **GR00T N1** [Bjorck *et al.*, 2025] uses latent actions as pseudo-targets for action decoding on human demonstrations, while jointly training with real action labels from heterogeneous robot data.

Nevertheless, despite their effectiveness in extracting knowledge from human videos, it is still an open question whether highly compressed latent tokens can sufficiently represent complex, high-DoF dexterous manipulation.

3.2 World Models

World-model-based approaches predict how the environment evolves over time. For human videos, this encompasses not only the future actions of the agent but also changes induced by its interactions with the environment. The underlying assumption is that accurately forecasting such dynamics equips the model with the action-relevant information necessary for task execution. Formally, this process can be expressed as $p(S_{t+1:t+H} \mid o_{\leq t}, l)$, where $S_{t+1:t+H}$ denotes future environment state (*e.g.*, raw pixels or visual features), conditioned on past observations $o_{\leq t}$ and language instruction l[1]. This formulation can be learned via video generation without action supervision, analogous to latent actions but more expressive for complex dynamics.

Existing methods mainly study world-model architectures and how to transfer their learned knowledge to VLA action prediction. As an early representative, **GR-1** [Wu *et al.*, 2024] proposes a GPT-style world model that autoregressively predicts future frames conditioned on language instructions and visual observations. The model is pretrained on egocentric human videos, and subsequently fine-tuned on robot datasets with an additional next-action prediction objective, leveraging internal features learned during frame generation. This paradigm is scaled by **GR-2** [Cheang *et al.*, 2024] using web-scale human data [Miech *et al.*, 2019] and a VQGAN-based visual tokens to improve the fidelity of future frame prediction, thereby facilitating the acquisition of more accurate action-related knowledge. **FLARE** [Zheng *et al.*, 2025] jointly trains its action decoder to perform action prediction and implicit world modeling by enforcing representation alignment (REPA) [Yu *et al.*, 2024] between decoder's internal features and future visual features on action-free hu-

[1]A stricter variant further conditions on future actions; in this survey, we focus on approaches that do not rely on action inputs.

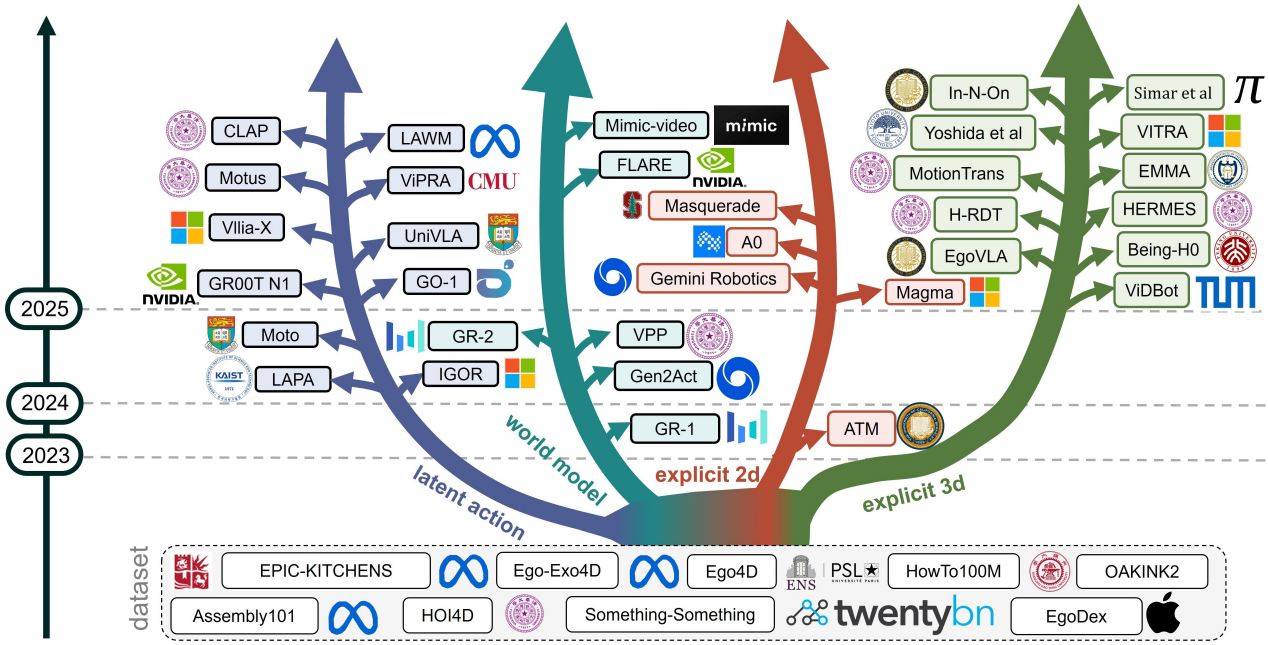

Figure 2: **Taxonomy of Human-to-Robot Policy Transfer.** Representative methods are organized into four paradigms by intermediate representations: **Latent Actions**, **World Models**, **Explicit 2D Representations**, and **Explicit 3D Representations**. The bottom row lists commonly used large-scale human-video datasets.

man videos. In contrast, **VPP** [Hu *et al.*, 2024] treats a pretrained video generation model (*i.e.*, SVD) as a feature extractor to extracts "predictive visual representations" as conditions for downstream action policies. More recently, **Mimic-Video** [Pai *et al.*, 2025] also shows that action prediction can be enhanced by leveraging intermediate representations from partially denoised steps of a pretrained video diffusion model as conditioning signals.

A complementary direction uses video generation as an explicit planning interface. **Gen2Act** [Bharadhwaj *et al.*, 2025] employs a pretrained video generation model to synthesize a "hallucinated" video plan from a static image. It constructs paired data by conditioning the generator on the initial frame of each robot demonstration to produce a corresponding human-like video, and then trains a video-conditioned policy that maps generated plans to robot actions.

Despite their promising results, world-model-based methods often incur substantial training overhead, since learning world models requires generating large numbers of visual tokens. In addition, effectively distilling and transferring the knowledge embedded in world modeling to action prediction remains a non-trivial and under-explored challenge.

3.3 Explicit 2D Representations

Some works instead rely on 2D visual primitives (*e.g.*, keypoints, bounding boxes) as intermediate representation. Unlike self-supervised latent actions or world models, these approaches leverage explicit and interpretable spatiotemporal cues, often extracted using 2D vision models, to bridge perception and action learning. The extracted cues are typically used as additional supervision to facilitate the acquisition of action-relevant knowledge from human videos.

For example, **ATM** [Wen *et al.*, 2024] uses dense point trajectories as action-like supervision, training a transformer to predict future motion that serves as a dynamic spatial prior for downstream manipulation. **Magma** [Yang *et al.*, 2025a] turns trajectories into an in-context control cue by rendering them as "Trace-of-Mark" visual prompts, effectively converting unlabelled videos into VLA-style training pairs with explicit temporal structure. At larger scale, **Gemini Robotics** [Team, 2025] trains an embodied reasoning backbone on human videos annotated with diverse 2D primitives (boxes, keypoints, and tracks), leveraging these signals to improve long-horizon generalization by incorporating these via techniques such as embodied chain-of-thought prompting [Zawalski *et al.*, 2024].

Beyond generic motion cues, **A0** [Xu *et al.*, 2025] focuses on explicit spatial affordance supervision by modeling object-centric contact points and post-contact trajectories as embodiment-agnostic interaction representations, enabling robust spatial reasoning across heterogeneous embodiments. From a complementary perspective, **Masquerade** [Lepert *et al.*, 2025] reduces the human–robot embodiment gap by editing in-the-wild human videos to visually resemble robot demonstrations, providing explicit 2D robot keypoint supervision that improves data efficiency and long-horizon generalization.

While these methods efficiently exploit existing vision tools to extract action-relevant cues, they may inherit label noise from the underlying models and often require additional filtering. Furthermore, 2D cues can be limited in expressing complex actions due to the lack of depth information, as real-world actions are inherently three-dimensional.

Paper	Source	Clips / Hours	Representation	End Effector
Latent Actions				
LAPA [2025]	SSv2	220k / –	VQ Latent Actions	Gripper
IGOR [2024]	SSv2, Epic, Ego4D, EGTEA	2m / –	VQ Latent Actions	Gripper
Moto [2025c]	SSv2	105k / –	VQ Latent Actions	Gripper
GO-1 [2025]	Ego4D	– / –	VQ Latent Actions	Gripper, Dex. Hand
GR00T [2025]	Ego4D, Ego-Exo4D, Assembly-101, Epic, HOI4D, HoloAssist, RH20T-Human	– / 2.5k	VQ Latent Actions	Gripper, Dex. Hand
UniVLA [2025]	Ego4D	– / –	VQ Latent Actions	Gripper
Villa-X [2025b]	Ego4D, EgoPAT3D, EGTEA Gaze+, Epic, HO-Cap, HOI4D, HoloAssist, RH20T, SSv2	3.6m / –	VQ Latent Actions	Gripper, Dex. Hand
ViPRA [2025]	SSv2	198k / –	VQ Latent Actions	Gripper
Motus [2025a]	Egodex	230k / –	Optical-Flow Latent Actions	Gripper
CLAP [2026]	Ego4D	– / 90h	VQ Latent Actions	Gripper
LAWM [2026]	YouTube Temporal-1B	– / 13k	Sparsity/Noise-Constrained Latent Actions	Gripper
World Models				
GR-1 [2024]	Ego4D	800k / 667	Generative Video Priors	Gripper
GR-2 [2024]	Ego4D, SSv2, Epic, HowTo100M, Kinetics-700	38m / –	Generative Video Priors	Gripper
VPP [2024]	SSv2	192k / –	Generative Video Priors	Gripper, Dex. Hand
FLARE [2025]	Self-collected	– / –	Future Visual Features	Gripper, Dex. Hand
mimic-video [2025]	Web-scale	– / –	Generative Video Priors	Gripper, Dex. Hand
Gen2Act [2025]	Web-scale	270m+ / –	Predicted Frames	Gripper
Explicit 2D				
ATM [2024]	Self-collected	0.3k / –	2D Point Trajectories	Gripper
Magma [2025a]	SSv2, Epic, Ego4D	4m / –	2D Point Trajectories	Gripper
Gemini Robotics [2025]	Web-scale	– / –	Boxes, Keypoints, Tracks	Gripper
A0 [2025]	HOI4D	22k / –	Contact Points/Trajectories	Gripper
Masquerade [2025]	Epic	675k* / –	2D Point Trajectories	Gripper
Explicit 3D				
VidBot [2025a]	Epic	– / –	3D Affordance	Gripper
EgoVLA [2025b]	HoloAssist, TACO, HOI4D, HOT3D	500k* / –	6D Wrist Pose/mano para.	Dex. Hand
Being-H0 [2025]	UniHand	2.5m / 1155	3D Hand Pose	Dex. Hand
H-RDT [2025b]	EgoDex	338k / 829h	6 Dof Wrist Pose/3D Fingertips	Gripper
MotionTrans [2025]	Self-collected	1.7k / –	6D Wrist Pose/Hand Joints	Dex. Hand
Yoshida et al. [2025]	Ego4D, EgoExo4D, Epic, Nymeria	45k / –	6-DoF Object Pose	Gripper
VITRA [2025]	Ego4D, Epic, EgoExo4D, SSv2	1m / –	3D Hand Pose	Dex. Hand
In-N-On [2025]	PH^2SD	– / 1k+	6 Dof Wrist Pose/3D Fingertips	Dex. Hand
Simar et al. [2025]	Self-collected	– / 14	3D Hand Pose	Gripper

Table 1: Taxonomy of representation bridges from human videos to robotic actions. We categorize existing methods based on the intermediate representation used to bridge the domain gap between human videos and robotic control. The table summarizes the video sources, data scale, specific representations, and the end-effectors used in the experiments for each method. Note: The symbol "*" indicates that the dataset size is measured in frames rather than video clips.

3.4 Explicit 3D Representations

Similar in spirit to 2D-based approaches, recent methods leverage 3D information extracted from human videos to train VLA policies. Compared to 2D cues, 3D representations provide a coordinate-consistent interface (*e.g.*, SE(3) trajectories, parametric hand states, or object poses), reducing viewpoint dependence and depth ambiguity, and aligning more naturally with VLA action prediction. In many systems, **parametric hand models** (*e.g.*, MANO [Romero *et al.*, 2017]) serve as a canonical and compact 3D representation of human manipulation and are widely used as a shared kinematic space linking human videos to robot control. Methods primarily vary in the types of videos they exploit, the strategies used to estimate 3D labels, and the mechanisms for bridging human and robot action spaces.

EgoVLA [Yang *et al.*, 2025b] leverages videos with high-quality 3D hand action annotations, obtained via multi-view optimization or RGB-D SLAM, to pretrain VLA models. It unifies human and robot action prediction by considering only wrist poses and finger tip positions, and adopts inverse kinematics from human action space for robot execution. **H-RDT** [Bi *et al.*, 2025b] adopts a similar hand-centric representation and leverages videos with hand annotations col-

lected using AR/VR devices. **In-N-On** [Cai *et al.*, 2025] also leverages AR-captured human videos and includes extra head poses into the unified action representation. **Being-H0** [Luo *et al.*, 2025] scales to larger human-video corpora and discretizes continuous MANO hand trajectories into a vocabulary of motion tokens for VLA pretraining.

Beyond learning from videos captured in controlled settings, recent methods have explored recovering 3D information from more in-the-wild videos. **VidBot** [Chen *et al.*, 2025a] combines structure-from-motion with depth foundation models to recover 3D affordance trajectories from unscripted videos. **VITRA** [Li *et al.*, 2025] proposes an automated pipeline that converts large collections of unscripted real-world human videos into robot-aligned atomic action segments with MANO hand annotations and language instructions, enabling more scalable pretraining. **Yoshida et al.** [Yoshida *et al.*, 2025] instead reconstruct 3D object pose changes from egocentric videos as supervisions.

In addition to the human-pretraining and robot-fine-tuning paradigm, recent work has investigated alternative learning schemes for VLA. For example, **MotionTrans** [Yuan *et al.*, 2025] and **Kareer et al.** [Kareer *et al.*, 2025] collect human demonstrations with 3D action annotations using wearable devices and jointly train VLA models on both human and robot data. They show that such co-training improves robots' generalization to previously unseen tasks.

Although 3D representations are better aligned with VLA action prediction and facilitate human-to-robot transfer, acquiring accurate 3D labels remains difficult. Wearable devices can provide high-quality annotations in controlled settings, but existing web-scale data must rely on 3D reconstruction from vision models, which is particularly challenging in in-the-wild scenarios and often introduces larger errors than 2D supervision. Moreover, directly predicting actions in 3D space requires models to not only understand image-space observations but also infer the underlying 3D world structure from 2D projections. From a generalization perspective, whether this paradigm is ultimately more effective than predicting in 2D image space remains an open question.

4 Human Video Datasets for VLA Learning

In this section, we provide a detailed review of the representative human video datasets underpinning the aforementioned approaches. We organize them along two axes, namely whether they provide *explicit metric 3D action labels* and whether collection is *scripted* or *unscripted*. Explicit 3D action labels determine whether the resulting representations can be directly grounded into robot-executable actions or must be inferred from visual state transitions [Blank *et al.*, 2024; Hirose *et al.*, 2024]. Table 2 summarizes this taxonomy.

Datasets without Explicit 3D Action Labels. These datasets provide RGB videos with semantic annotations (e.g., scenes, objects, action descriptions) but lack metric 3D trajectories of hands or bodies in a camera/world frame, such as 3D keypoints or MANO parameters. Therefore, they are mainly used to support *latent-action* or *world-model*-based approaches, where actions are inferred from visual state transitions rather than being grounded directly in a 3D action space.

Representative datasets include Something-Something V2 (short hand–object clips for manipulation dynamics) [Goyal *et al.*, 2017], EPIC-KITCHENS (large-scale egocentric kitchen activities) [Damen *et al.*, 2020], Ego4D (egocentric daily-life video across diverse scenarios) [Grauman *et al.*, 2022], HowTo100M (Internet-scale instructional video) [Miech *et al.*, 2019], Egocentric-100K (100K hours from factory environments, largely without annotations) [AI, 2025], and Ego-Exo4D (synchronized ego–exo multi-view recordings for cross-view learning) [Grauman *et al.*, 2024]. Despite their scale and diversity, they are typically used as video sources rather than as providers of metric 3D action representations. In terms of collection protocol, Something-Something V2 is scripted via predefined action templates [Goyal *et al.*, 2017], whereas EPIC-KITCHENS, Ego4D, Ego-Exo4D, HowTo100M, and Egocentric-100K are unscripted [Damen *et al.*, 2020; Grauman *et al.*, 2022; Grauman *et al.*, 2024; Miech *et al.*, 2019; AI, 2025]. The lack of predefined boundaries in continuous in-the-wild recordings complicates action segmentation and instruction annotation for imitation learning.

Datasets with Explicit 3D Action labels. A smaller but crucial class provides *explicit metric 3D action annotations*, including 3D keypoint trajectories of hands/bodies and parametric representations such as MANO or SMPL. They are typically captured with instrumented setups (e.g., AR/VR headsets, depth sensors, or motion capture) that record motion in camera/world coordinates. Representative datasets include EgoDex [Hoque *et al.*, 2025], HOI4D [Liu *et al.*, 2022], Assembly101 [Sener *et al.*, 2022], ARCTIC [Fan *et al.*, 2023], and OakInk2 [Zhan *et al.*, 2024]. HOI4D provides egocentric RGB-D with frame-wise 3D hand and object poses [Liu *et al.*, 2022]. EgoDex uses head-mounted AR to collect paired RGB video and 3D hand/finger trajectories over hundreds of tabletop tasks, targeting dexterous imitation learning [Hoque *et al.*, 2025]. Assembly101 focuses on long-horizon procedural activities with synchronized static and egocentric views, dense segmentation, and millions of 3D hand poses [Sener *et al.*, 2022]. ARCTIC captures bimanual object articulation with 3D hand/object meshes and dynamic contact information [Fan *et al.*, 2023]. OakInk2 adopts an object-centric hierarchy (affordances, primitives, complex tasks) and provides multi-view data with precise 3D annotations of hands, bodies, and objects for structured task decomposition and motion generation [Zhan *et al.*, 2024]. These datasets enable direct supervision with explicit 2D/3D action labels and are closest to robot imitation learning, often improving data efficiency [Li *et al.*, 2025]. However, they are usually smaller-scale and collected in laboratory or instrumented settings, trading diversity for annotation fidelity. Consistent with this, collection is commonly *scripted*, with predefined interactions/objects and carefully segmented action units that align well with imitation-learning supervision, but limit behavioral diversity, scalability, and representativeness relative to large unscripted corpora.

Dataset	Scale	View	Lang. Ann.	Geo Signal	Scripted
Category 1: Datasets w/o Explicit 3D Action Labels					
Something-Something V2 (SSv2) [Goyal *et al.*, 2017]	220K clips	3rd	✓	RGB	✓
EPIC-KITCHENS-100 [Damen *et al.*, 2020]	100 hours	Ego	✓	RGB	✗
Ego4D [Grauman *et al.*, 2022]	3.6K hours	Ego	✓	RGB	✗
EgoExo4D [Grauman *et al.*, 2024]	1.3K hours	MV + Ego	✓	MV (calib)	✗
HowTo100M [Miech *et al.*, 2019]	136M clips	Mix	✓	RGB	✗
Egocentric-100K [AI, 2025]	100K hours	Ego	✗	RGB	✗
Category 2: Datasets w/ Explicit 3D Action Labels					
HOI4D [Liu *et al.*, 2022]	2.4M frames	Ego	✗	RGB-D	✓
Assembly-101 [Sener *et al.*, 2022]	513K clips	MV + Ego	✓	MV (calib)	✓
EgoDex [Hoque *et al.*, 2025]	300K episodes	Ego	✓	RGB	✓
OAKINK2 [Zhan *et al.*, 2024]	4.0M frames	MV + Ego	✓	MV (calib)	✓
ARCTIC [Fan *et al.*, 2023]	2.1M frames	MV + Ego	✓	MV (calib)	✓

Table 2: Classification of human-video datasets for VLA. Scale reports dataset size using the most standard unit for each dataset. Geo Signal indicates available geometric/3D-related signals, not necessarily explicit 3D labels. Scripted denotes whether data collection follows predefined tasks/procedures rather than fully in-the-wild recording.

5 Challenges and Future Directions

In this section, we analyze the open challenges commonly faced by diverse approaches that leverage human videos, organizing them around three interfaces in the human-video-to-robot pipeline. These include episodizing unstructured human videos into training-ready episodes, handling heterogeneity when mapping video-derived supervision to robot-executable actions, and designing evaluations that predict real deployment performance. This decomposition isolates the main failure modes and clarifies the abstractions needed for reliable transfer.

5.1 Transforming Human Videos to Episodes

A bottleneck appears upstream of policy learning. Web-scale human video from platforms such as YouTube and TikTok is rarely packaged into episodes that match the supervision granularity required by robot learning. Instructional corpora such as HowTo100M [Miech *et al.*, 2019] offer a scalable proxy but suffer from weak narration–manipulation alignment, while egocentric datasets [Grauman *et al.*, 2022; Grauman *et al.*, 2024; Damen *et al.*, 2020] remain largely untrimmed and often contain viewpoint changes and off-task segments. As a result, clips frequently mix intents, and similar visual transitions can correspond to different latent actions. This ambiguity degrades learning signals by destabilizing discrete tokenization and diverting model capacity to control-irrelevant dynamics. Procedural datasets such as HOI4D and Assembly101 [Liu *et al.*, 2022; Sener *et al.*, 2022] mitigate this issue by aligning video units with manipulation primitives, whereas in-the-wild video requires explicit episodization, as explored by systems such as VITRA [Li *et al.*, 2025]. Future episodization should be semantic and interaction-driven, defining segments around manipulation phases and object state changes rather than fixed time windows. Standardized construction protocols are also needed to reduce hidden variance and support rigorous comparison.

5.2 Addressing Heterogeneity

Embodiment Mismatch. Embodiment mismatch stems from kinematic and contact-mechanical differences between human hands and robotic end-effectors. Retargeting is intrinsically under-constrained when high-DoF human motion is mapped onto lower-DoF robot control spaces, so distinct human motions can yield the same robot command and some motions admit no feasible realization [Qin *et al.*, 2022]. The issue is amplified for grippers and across diverse end-effector morphologies, where action spaces are not aligned and cross-morphology VLA training becomes difficult [Bi *et al.*, 2025b]. A practical response is to prioritize embodiment-invariant supervision, such as interaction intent and object-centric effects, rather than literal trajectory imitation. Hardware advances may reduce the gap [Shenzhen ShARPA Robotics, 2024; Tesla, Inc., 2022], but reliable transfer still requires morphology-conditioned grounding that maps shared priors into feasible commands under platform-specific constraints.

Viewpoint Heterogeneity. Robotic manipulation is typically trained and deployed with head-mounted observations as the primary view, often supplemented by wrist or eye-in-hand cameras for contact-centric detail [Black *et al.*, 2024]. Web-scale human video, however, is predominantly exocentric, where fine contacts are weakly observed and occlusions around the interaction site are common [Miech *et al.*, 2019]. Egocentric human video is closer but still mismatched, since head-mounted streams (e.g., Ego4D, EPIC-KITCHENS) exhibit rapid viewpoint shifts driven by head turns and frequent occlusions, which can remove the active hand–object contact from view during manipulation [Grauman *et al.*, 2022; Damen *et al.*, 2020]. Progress likely requires explicit cross-view alignment between human and robot observations and an action representation that is anchored to interaction outcomes, for example object state changes, rather than view-specific appearance.

5.3 Benchmarking and Evaluation

A key challenge is that existing benchmarks only partially reflect the regime in which human video supervision is intended to help. Standard suites such as LIBERO, CALVIN, and SIMPLER support controlled comparison, but their task templates and simulated scenes under-represent the diversity and long-horizon structure typical of Internet human videos [Liu *et al.*, 2023; Mees *et al.*, 2022; Li *et al.*, 2024b]. As a result, improvements can be benchmark-specific and do not directly answer the central question, namely transfer efficiency under a fixed robot-data budget. Real-robot evaluation remains necessary, yet scaling and standardizing it is difficult due to cost, safety, and platform variability. Progress requires evaluation protocols that track deployment-relevant generalization and transfer efficiency. This includes reporting performance as a function of robot-data budget, using held-out splits over novel objects and scenes, and quantifying robustness under viewpoint and embodiment shifts. Testbeds closer to human activity distributions such as BEHAVIOR-1K [Li *et al.*, 2024a], together with scalable digital-twin or world-model toolchains such as Cosmos [Agarwal *et al.*, 2025], can increase coverage and reproducibility, making results more interpretable and comparable.

6 Conclusion

Human video offers a scalable route to easing the data bottleneck in VLA learning, but only if human observation can be bridged to robot execution. This survey organized recent progress around representation bridges that infer latent actions, learn predictive world models, or construct explicit 2D/3D structure. These paradigms offer complementary strengths: implicit representations scale across diverse video sources, while explicit geometric cues provide stronger grounding and clearer interfaces to robot-executable actions. Future progress will depend on episodizing in-the-wild video, handling kinematic and viewpoint mismatch, and developing evaluation protocols that better predict deployment performance and transfer efficiency.

References

[Agarwal *et al.*, 2025] Niket Agarwal, Arslan Ali, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[AgiBot-World-Contributors, 2025] AgiBot-World-Contributors. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *IROS*, pages 3549–3556, 2025.

[AI, 2025] Build AI. Egocentric-100k, 2025.

[Bahl *et al.*, 2023] Shikhar Bahl, Russell Mendonca, et al. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.

[Bharadhwaj *et al.*, 2025] Homanga Bharadhwaj, Debidatta Dwibedi, et al. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In *CoRL*, volume 305 of *PMLR*, pages 3936–3951. PMLR, 2025.

[Bi *et al.*, 2025a] Hongzhe Bi, Hengkai Tan, et al. Motus: A unified latent action world model. *arXiv:2512.13030*, 2025.

[Bi *et al.*, 2025b] Hongzhe Bi, Lingxuan Wu, et al. H-RDT: Human manipulation enhanced bimanual robotic manipulation. *arXiv:2507.23523*, 2025.

[Bjorck *et al.*, 2025] Johan Bjorck, Fernando Castañeda, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv:2503.14734*, 2025.

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

[Blank *et al.*, 2024] Nils Blank, Moritz Reuss, et al. Scaling robot policy learning via zero-shot labeling with foundation models. *Conference on Robot Learning (CoRL)*, 2024.

[Borja-Diaz *et al.*, 2022] Jessica Borja-Diaz, Oier Mees, et al. Affordance learning from play for sample-efficient policy learning. In *ICRA*, Philadelphia, USA, 2022.

[Brohan *et al.*, 2023] Anthony Brohan, Noah Brown, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. arXiv preprint available.

[Bu *et al.*, 2025] Qingwen Bu, Yanting Yang, et al. UniVLA: Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[Cai *et al.*, 2025] Xiongyi Cai, Ri-Zhao Qiu, et al. In-N-On: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv:2511.15704*, 2025.

[Cheang *et al.*, 2024] Chi-Lam Cheang, Guangzeng Chen, et al. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv:2410.06158*, 2024.

[Chen *et al.*, 2024] Xiaoyu Chen, Junliang Guo, et al. Igor: Image-goal representations are the atomic control units for foundation models in embodied AI. *arXiv:2411.00785*, 2024.

[Chen *et al.*, 2025a] Hanzhi Chen, Boyang Sun, et al. VidBot: Learning generalizable 3D actions from in-the-wild 2D human videos. In *CVPR*, 2025.

[Chen *et al.*, 2025b] Xiaoyu Chen, Hangxing Wei, et al. Villa-x: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025.

[Chen *et al.*, 2025c] Yi Chen, Yuying Ge, et al. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *ICCV*, 2025.

[Damen *et al.*, 2020] Dima Damen, Hazel Doughty, et al. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *TPAMI*, 43(11), 2020.

[Fan *et al.*, 2023] Zicong Fan, Omid Taheri, et al. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, pages 12943–12954, 2023.

[Garrido *et al.*, 2026] Quentin Garrido, Tushar Nagarajan, et al. Learning latent action world models in the wild, 2026.

[Goyal *et al.*, 2017] Raghav Goyal, Samira Ebrahimi Kahou, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[Grauman *et al.*, 2022] Kristen Grauman, Andrew Westbury, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[Grauman *et al.*, 2024] Kristen Grauman, Andrew Westbury, et al. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

[Hirose *et al.*, 2024] Noriaki Hirose, Catherine Glossop, et al. Lelan: Learning a language-conditioned navigation policy from in-the-wild videos. In *Conference on Robot Learning*, 2024.

[Hoque *et al.*, 2025] Ryan Hoque, Peide Huang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[Hu *et al.*, 2024] Yucheng Hu, Yanjiang Guo, et al. Video prediction policy: A generalist robot policy with predictive visual representations. In *ICML*. PMLR, 2024.

[Kareer *et al.*, 2025] Simar Kareer, Karl Pertsch, et al. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025.

[Khazatsky *et al.*, 2024] Alexander Khazatsky, Karl Pertsch, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *DGR@RSS 2024 (Poster)*, 2024.

[Kim *et al.*, 2024] Moo Jin Kim, Karl Pertsch, et al. OpenVLA: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[Lepert *et al.*, 2025] Marion Lepert, Jiaying Fang, et al. Masquerade: Learning from in-the-wild human videos using data-editing. In *H2R Workshop at the Conference on Robot Learning (CoRL)*, 2025.

[Li *et al.*, 2024a] Chengshu Li, Ruohan Zhang, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

[Li *et al.*, 2024b] Xuanlin Li, Kyle Hsu, et al. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning*, 2024.

[Li *et al.*, 2025] Qixiu Li, Yu Deng, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv:2510.21571*, 2025.

[Liu *et al.*, 2022] Yunze Liu, Yun Liu, et al. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022.

[Liu *et al.*, 2023] Bo Liu, Yifeng Zhu, et al. Libero: Benchmarking knowledge transfer for lifelong robot learning. *NIPS*, 36:44776–44791, 2023.

[Luo *et al.*, 2025] Hao Luo, Yicheng Feng, et al. Being-H0: Vision-language-action pretraining from large-scale human videos. *arXiv:2507.15597*, 2025.

[Ma *et al.*, 2023] Jason Yecheng Ma, Shagun Sodhani, et al. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023. Spotlight presentation.

[Mandikal and Grauman, 2022] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022.

[Mees *et al.*, 2022] Oier Mees, Lukas Hermann, et al. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[Mees *et al.*, 2023] Oier Mees, Jessica Borja-Diaz, et al. Grounding language with visual affordances over unstructured data. In *ICRA*, London, UK, 2023.

[Miech *et al.*, 2019] Antoine Miech, Dimitri Zhukov, et al. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[Nair *et al.*, 2022] Suraj Nair, Aravind Rajeswaran, et al. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[O'Neill *et al.*, 2024] Abby O'Neill, Abdul Rehman, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *ICRA*, 2024.

[Pai *et al.*, 2025] Jonas Pai, Liam Achenbach, et al. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

[Qin *et al.*, 2022] Yuzhe Qin, Yueh-Hua Wu, et al. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, 2022.

[Romero *et al.*, 2017] Javier Romero, Dimitrios Tzionas, et al. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

[Routray *et al.*, 2025] Sandeep Routray, Hengkai Pan, et al. ViPRA: Video prediction for robot actions. *arXiv:2511.07732*, 2025.

[Sener *et al.*, 2022] Fadime Sener, Dibyadip Chatterjee, et al. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, 2022.

[Shenzhen ShARPA Robotics, 2024] Shenzhen ShARPA Robotics. Wave humanoid robot. https://www.sharpa.com/pages/wave, 2024. Accessed 2025-01.

[Tan *et al.*, 2025] Reuben Tan, Baolin Peng, et al. Multimodal reinforcement learning with agentic verifier for ai agents. *arXiv preprint arXiv:2512.03438*, 2025.

[Team, 2025] Gemini Robotics Team. Gemini Robotics: Bringing AI into the physical world. *arXiv:2503.20020*, 2025.

[Tesla, Inc., 2022] Tesla, Inc. Tesla ai day 2022. https://www.youtube.com/watch?v=ODSJsviD_SU, September 2022. Official technical presentation.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NIPS*, 30, 2017.

[Wen *et al.*, 2024] Chuan Wen, Xingyu Lin, et al. Any-point trajectory modeling for policy learning. In *RSS*, 2024.

[Wu *et al.*, 2024] Hongtao Wu, Ya Jing, et al. Unleashing large-scale video generative pre-training for visual robot manipulation. In *ICLR*, 2024. Poster presentation.

[Xiao *et al.*, 2022] Tete Xiao, Ilija Radosavovic, et al. Masked visual pre-training for motor control. *arXiv:2203.06173*, 2022.

[Xu *et al.*, 2025] Rongtao Xu, Jian Zhang, et al. A0: An affordance-aware hierarchical model for general robotic manipulation. In *ICCV*, 2025.

[Yang *et al.*, 2025a] Jianwei Yang, Reuben Tan, et al. Magma: A foundation model for multimodal AI agents. In *CVPR*, 2025.

[Yang *et al.*, 2025b] Ruihan Yang, Qinxi Yu, et al. EgoVLA: Learning vision-language-action models from egocentric human videos. *arXiv:2507.12440*, 2025.

[Ye *et al.*, 2025] Seonghyeon Ye, Joel Jang, et al. Latent action pretraining from videos. In *ICLR*, 2025. Poster presentation.

[Yoshida *et al.*, 2025] Tomoya Yoshida, Shuhei Kurita, et al. Developing vision-language-action model from egocentric videos. *arXiv:2509.21986*, 2025.

[Yu *et al.*, 2024] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, et al. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[Yuan *et al.*, 2025] Chengbo Yuan, Rui Zhou, et al. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. In *H2R Workshop at the Conference on Robot Learning (CoRL)*, 2025.

[Zawalski *et al.*, 2024] Michał Zawalski, William Chen, et al. Robotic control via embodied chain-of-thought reasoning. In *Conference on Robot Learning*, 2024.

[Zhan *et al.*, 2024] Xinyu Zhan, Lixin Yang, et al. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *CVPR*, pages 445–456, 2024.

[Zhang *et al.*, 2026] Chubin Zhang, Jianan Wang, et al. Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos, 2026.

[Zheng *et al.*, 2025] Ruijie Zheng, Jing Wang, et al. FLARE: Robot learning with implicit world modeling. In *SWOMO Workshop at Robotics: Science and Systems (RSS)*, 2025.