VISION-LANGUAGE-ACTION MODELS FOR ROBOTICS: A REVIEW TOWARDS REAL-WORLD APPLICATIONS

Kento Kawaharazuka[a,*], **Jihoon Oh**[a], **Jun Yamada**[b], **Ingmar Posner**[b,†], **Yuke Zhu**[c,†]
[a] *University of Tokyo*
[b] *University of Oxford*
[c] *The University of Texas at Austin*

ABSTRACT

Amid growing efforts to leverage advances in large language models (LLMs) and vision-language models (VLMs) for robotics, Vision-Language-Action (VLA) models have recently gained significant attention. By unifying vision, language, and action data at scale, which have traditionally been studied separately, VLA models aim to learn policies that generalise across diverse tasks, objects, embodiments, and environments. This generalisation capability is expected to enable robots to solve novel downstream tasks with minimal or no additional task-specific data, facilitating more flexible and scalable real-world deployment. Unlike previous surveys that focus narrowly on action representations or high-level model architectures, this work offers a comprehensive, full-stack review, integrating both software and hardware components of VLA systems. In particular, this paper provides a systematic review of VLAs, covering their strategy and architectural transition, architectures and building blocks, modality-specific processing techniques, and learning paradigms. In addition, to support the deployment of VLAs in real-world robotic applications, we also review commonly used robot platforms, data collection strategies, publicly available datasets, data augmentation methods, and evaluation benchmarks. Throughout this comprehensive survey, this paper aims to offer practical guidance for the robotics community in applying VLAs to real-world robotic systems. All references categorized by training approach, evaluation method, modality, and dataset are available in the table on our project website: https://vla-survey.github.io.

1 Introduction

The recent success in developing a variety of large language models (LLMs) [1, 2] and large vision-language models (VLMs) [3, 4] has catalised remarkable advances in natural language processing and computer vision, fundamentally transforming both fields. These advancements have also been extended to the field of robotics, where LLMs and VLMs are leveraged to interpret multimodal inputs, reason about tasks, and perform context-aware actions, thereby laying the groundwork for more generalisable and scalable robotic systems [5–7].

Earlier works decouple LLMs and VLMs from the underlying robot policies responsible for action generation [8, 9]. While effective for a limited set of predefined tasks, such systems typically rely on selecting from fixed motion primitives or on policies learned through imitation learning, which limits their ability to generalise to a broader range of tasks. Learning policies that can generalise from current observations and instructions to unseen tasks remains a significant challenge.

To overcome these limitations, a growing body of research focuses on Vision-Language-Action (VLA) models [10]. By jointly learning visual, linguistic, and action modalities in an end-to-end framework, VLAs aim to enable robots to perform a wider range of tasks. The hope is that the resulting generalist policies aim to achieve generalization across diverse tasks and facilitate effective transfer across varying robotic embodiments. This approach reduces the need for extensive task-specific data collection and training, significantly lowering the cost of real-world deployment. As such, VLAs offer a promising path toward more scalable and accessible robotic systems.

[*] Corresponding author. Email: kawaharazuka@jsk.imi.i.u-tokyo.ac.jp

[†] Equal contribution.

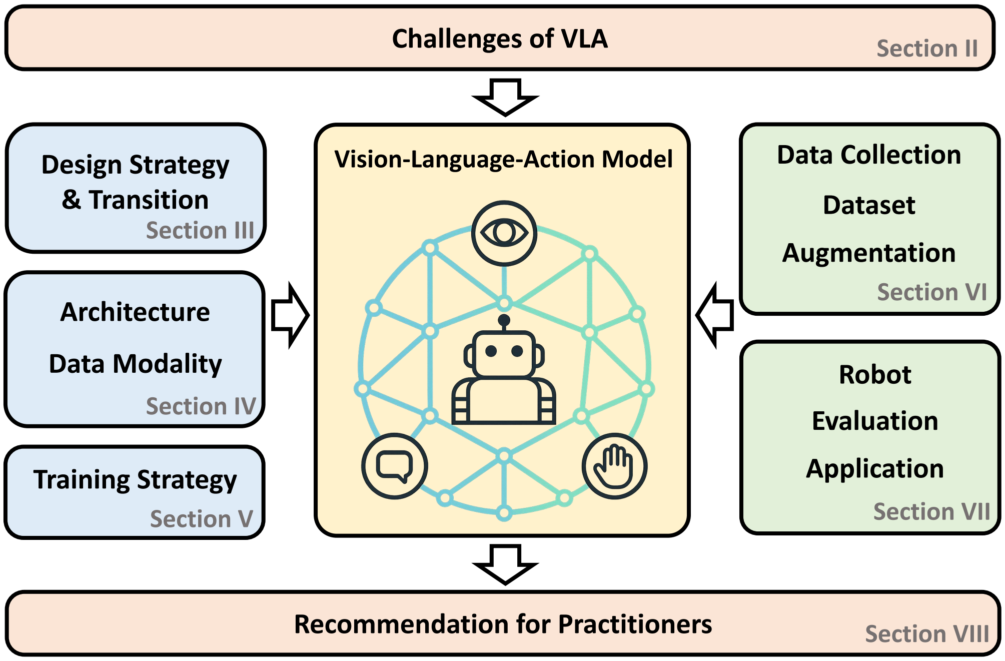


Figure 1: **Structure of this survey.** Section 2 outlines the key challenges in developing Vision-Language-Action (VLA) models. Section 3 and Section 4 review the evolution of VLA strategies, architectures, and modality-specific design choices. Section 5 categorizes training strategies and practical implementation considerations. Section 6 discusses the data collection methologies, publicly available dataset, and data augmentation. Section 7 discusses the types of robots used, evaluation benchmarks, and the applications of VLA models in real-world robot systems. Guidance for practitioners is presented in Section 8, based on the findings of the systematic review.

Despite growing interest, research on VLAs remains in its early stages. Architectural and training methodologies are not yet standardized, making it difficult to form a cohesive understanding of the field. This survey provides a systematic overview of the current landscape of VLAs, including their historical development, model architectures, modality integration strategies, and learning paradigms. While several previous surveys [11–13] have focused primarily on either action tokenization or general architectural advancements, this survey provides a comprehensive, full-stack overview, covering both software and hardware components. Specifically, beyond architecture and the development of VLAs, it includes robot platforms, data collection strategies, publicly available datasets, data augmentation techniques, and evaluation benchmarks. We also introduce a taxonomy of existing VLA models and analyze representative models within each category. This survey is intended to serve as a practical guide for researchers aiming to apply VLA models to real-world robotic systems.

In this review, to clarify the scope, we define VLA models as systems that take visual observations and natural language instructions as core inputs and produce robot actions by directly generating control commands (see Def. 1). While additional modalities (e.g., proprioception or depth) may be included, the integration of vision and language is essential. We exclude approaches that use vision and language solely for high-level reasoning or task planning without grounding them in action execution, such as those that select from a set of pre-trained skills using a high-level policy.

Definition 1.1

A Vision-Language-Action (VLA) model is a system that takes visual observations and natural language instructions as required inputs and may incorporate additional sensory modalities. It produces robot actions by directly generating control commands. Thus, models in which a high-level policy (e.g., a vision-language model backbone) merely selects an index from a set of pre-trained skills or control primitives are excluded from this definition.

The overall structure of this survey is illustrated in Fig. 1. First, Section 2 outlines the key challenges addressed in VLA research. Section 3 reviews major strategies and the architectural transition of VLA models. Section 4 introduces core architectural components and building blocks, including modality-specific processing modules. Section 5 discusses key training strategies and practical implementation considerations. Section 6 summarises data collection methodologies, publicly available datasets, and data augmentation. Then, Section 7 provides guidance for real-world deployment, covering commonly used robot platforms, evaluation protocols, and current real-world applications. Based on the findings of the systematic review, we present recommendations for practitioners in Section 8. Finally, Section 9 discusses open challenges and future directions, and Section 10 presents our concluding remarks.

2 Challenges

The integration of visual, linguistic, and motor modalities presents a promising pathway toward the development of generalist robot policies. However, the advancement of robust and deployable VLA models is still constrained by several fundamental challenges. These limitations span across data availability, embodiment mismatches, and computational constraints, each imposing critical design trade-offs in model architecture, training strategy, and deployment feasibility.

2.1 Data Requirements and Scarcity

Training VLA models require large-scale, diverse, and well-annotated data that aligns visual observations with natural language instructions and corresponding actions. However, datasets satisfying all three modalities, vision, language, and action, are limited in both scale and diversity. While vision-language datasets such as COCO Captions [14] or web-scale corpora offer broad linguistic grounding, they lack the action grounding necessary for robotics. Conversely, robot demonstration datasets often contain limited linguistic variability or are confined to narrow task distribution.

This mismatch leads to two data-related bottlenecks. First, models pre-trained on large-scale web or video datasets may not transfer effectively to robotic tasks due to a lack of motor grounding or a discrepancy in the domain. Second, high-quality robot demonstrations, often collected via teleoperation are expensive and difficult to scale. Such an issue is further exacerbated when the number of modalities increases, such as adding tactile, acoustic, and 3D information.

2.2 Embodiment Transfer

Robots exhibit a wide range of embodiments. Some are equipped solely with arms, while others incorporate wheels, legs, or other mobility mechanisms. Their joint configurations, link structures, sensor types and placements, and even physical appearances vary significantly. While VLA models are increasingly trained on data from diverse robot embodiments, transferring policies across embodiments remains a major challenge. Each robot typically operates in a distinct action space and proprioceptive observation space, reflecting differences in degrees of freedom, sensor modalities, and kinematic structure.

A related challenge lies in leveraging human motion data for training. Given the high cost of collecting large-scale robot data, human demonstrations offer a promising alternative. However, such data generally lack explicit action labels, and even when actions are inferred, they differ substantially from robot actions in both form and semantics. As with robot-to-robot transfer, mapping human demonstrations into robot-executable actions is highly non-trivial.

These embodiment-related challenges raise fundamental questions for VLA development: What kinds of data best support cross-embodiment generalization? How should morphological and sensory differences be represented? And how can models be trained to ensure robust grounding of vision and language across diverse robotic and human embodiments?

2.3 Computational and Training Cost

Training VLA models entails a considerable amount of computational demands due to the high-dimensional and multi-modal nature of their input, typically including vision, language, and actions. While many recent approaches leverage pre-trained VLM as a backbone, these models are typically adapted for robotics domain via large-scale robot demonstrations or simulated data. Most practitioners are expected to build upon such pre-trained models and further fine-tune them for downstream tasks using task-specific, high-quality expert demonstrations, rather than training end-to-end from scratch. Nonetheless, both the adaptation and fine-tuning stages remain computationally intensive, especially when processing long temporal sequences, high-resolution images, or additional modalities such as 3D point clouds or proprioception. Transformer-based architectures, which dominate current VLA designs, also scale poorly with respect to sequence length and input dimensionality, further amplifying memory and compute costs. At inference time, running these models in real-world settings, particularly on resource-constrained robotic platforms, poses additional challenges related to latency and memory usage. These computational burdens limit the accessibility and deployability of VLA systems, motivating ongoing research into efficient model architectures and distillation methods that can reduce resource requirements without significantly degrading performance.

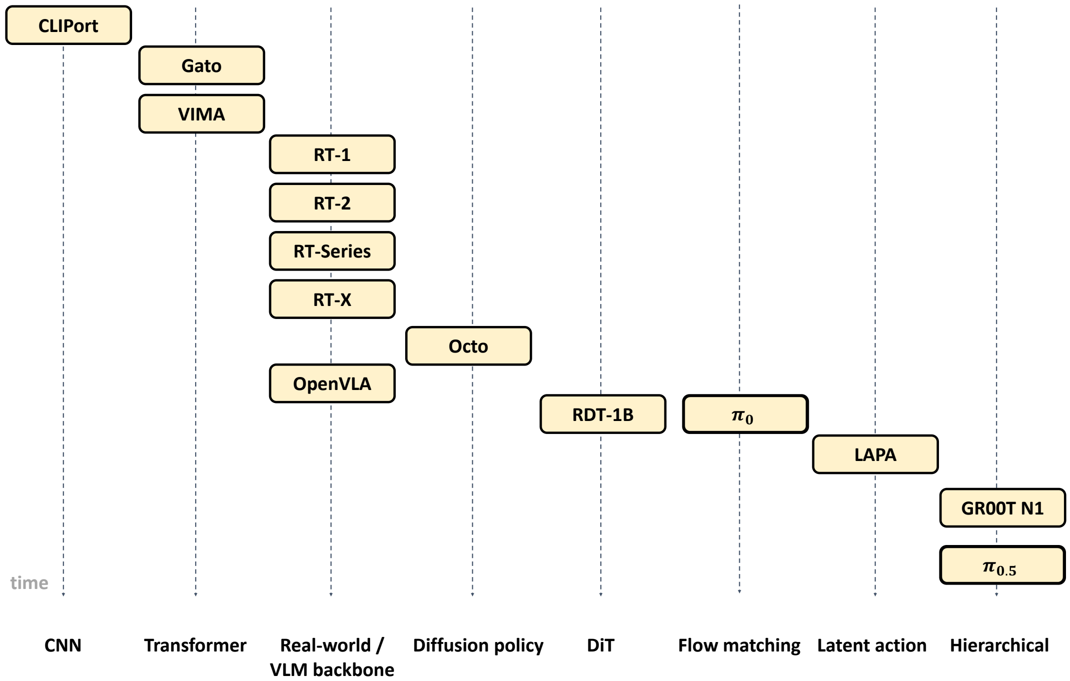


Figure 2: **Timeline of major Vision–Language–Action (VLA) models.** This figure summarizes the historical progression of representative VLA models shown in Section 3: from early CNN-based models (e.g., *CLIPort* [15]), to real-world scalable policies leveraging pre-trained VLM backbones (e.g., *RT-1*, *RT-2*, *RT-X*, *OpenVLA* [10, 16–18]), followed by models integrating diffusion and flow matching techniques (e.g., *Octo*, *RDT-1B*, π_0 [19–21]), and more recent approaches focusing on latent action inference and hierarchical control (e.g., *LAPA*, $\pi_{0.5}$, *GR00T N1* [22–24]).

3 VLA Design Strategy and Transition

This section categorizes major interface strategies for transforming vision and language inputs into robot actions, following the historical progression of VLA architectures (see Fig. 2). Each architectural category corresponds to a distinct generation of VLA systems, characterized by how multimodal representations are aligned with control. The discussion spans from early CNN-based models to transformer-based architectures, diffusion-based policies, and finally, hierarchical control frameworks.

Early CNN-based end-to-end architectures. A foundational approach to end-to-end VLAs is CLIPort [15], one of the earliest models to integrate CLIP [25] for extracting visual and linguistic features. It combines these modalities with the Transporter Network [26] to learn object manipulation tasks in an end-to-end manner, identifying which object to move and where to place it. CLIPort demonstrated the feasibility of jointly training vision, language, and action by leveraging CLIP [25] as a pre-trained VLM. However, approaches based on Convolutional Neural Networks (CNNs) and Multi-Layer Perceptrons (MLPs) face challenges in unifying diverse modalities and also struggle to scale effectively.

Transformer-based sequence models. To address these limitations, Google DeepMind released Gato [27], a generalist agent and precursor to the Robotics Transformer (RT) series. Gato performs a wide range of tasks, such as text chatting, visual question answering, image captioning, gameplay, and robot control, using a single transformer [28] model. It tokenizes language instructions using SentencePiece [29] and encodes images using Vision Transformer (ViT) [30]. A decoder-only transformer is then used to autoregressively generate actions based on the combined input sequence. While Gato enables multiple tasks with a single network, its repertoire of robotic skills remains limited to a narrow set, such as block stacking with a robotic arm. Similarly, VIMA [31] is an encoder-decoder transformer model that enables robots to follow general task instructions provided through a combination of text and goal images. Objects are first detected using Mask R-CNN [32], after which each detected object's image is tokenized using ViT.

Bounding box coordinates are separately embedded as tokens, and textual instructions are tokenized using the T5 tokenizer [33]. A frozen T5 encoder and a transformer decoder are then used to autoregressively generate discrete action tokens. While VIMA demonstrates the ability to perform a wide range of robotic tasks, all experiments were limited to simulation environments.

Unified real-world policies with pre-trained VLMs. To enable scalable real-world applications, Robotics Transformer-1 (RT-1) [16] has been introduced as a real-time, general-purpose control model capable of performing a wide range of real-world tasks. RT-1 processes a sequence of images using EfficientNet [34], and performs FiLM conditioning [35] with language features encoded by the Universal Sentence Encoder (USE) [36], enabling early fusion of visual and linguistic modalities. The extracted tokens are compressed via TokenLearner [37] and then passed through a decoder-only transformer, which outputs discretized action tokens nonautoregressively (see Section 4.1). Trained on a large-scale dataset comprising 700 tasks and 130,000 episodes, RT-1 is regarded as the first VLA that unifies a broad range of robotic tasks. Subsequently, RT-2 [10] has been introduced as the successor to RT-1. It builds on a Vision-Language Model (VLM) backbone such as PaLM-E [38] or PaLI-X [39], pre-trained on large-scale internet data. RT-2 is jointly fine-tuned on both internet-scale vision-language tasks and robotic data from RT-1, resulting in strong generalization to novel environments. This VLM-based design has since become the standard architecture for VLAs. In contrast, RT-X [17] has been introduced to demonstrate that training on datasets collected from multiple robots enables the development of more general-purpose VLAs, moving beyond the single-robot training paradigm of RT-1 and RT-2.

The RT series has been extended into several variations, including RT-Sketch, which takes sketch images as input; RT-Trajectory, which takes motion trajectories as input; and others such as RT-H, Sara-RT, and AutoRT [40–44]. Among these, RT-H [42] is particularly notable for introducing a hierarchical policy structure. Built on the RT-2 architecture, RT-H incorporates a high-level policy that predicts an intermediate representation known as language motion, and a low-level policy that generates actions based on it. By modifying the input prompt, the model can flexibly alternate between generating high-level actions expressed in language and producing low-level robot actions directly. By sequentially switching between high-level and low-level policies, RT-H demonstrates improved performance, particularly in long-horizon tasks. Such hierarchical VLA architectures have since become a recurring design pattern in subsequent models. Building upon the RT-series, OpenVLA [18] is introduced as an open-source VLA framework that closely mirrors the architecture of RT-2, leveraging a pre-trained VLM as its backbone. Specifically, it employs Prismatic VLM [45], based on LLaMa 2 (7B) [1], and encodes image inputs using DINOv2 [46] and SigLIP [47]. Through full fine-tuning on the Open-X Embodiment (OXE) dataset [17], OpenVLA outperforms both RT-2 and Octo, and has since emerged as a mainstream architecture for VLA.

Diffusion policy. Octo [19], introduced after the RT series, is the first VLA to leverage Diffusion Policy [48], and also gained attention for its fully open-source implementation. Octo supports flexible goal specification, which can include a language instruction and a goal image, processed by a T5 encoder and a CNN, respectively. For input observations, it similarly uses a CNN to encode images and a lightweight multilayer perceptron (MLP) to embed proprioceptive signals. All tokens are concatenated into a single sequence, augmented with modality-specific learnable tokens, and passed into a transformer. Finally, a diffusion policy generates continuous actions, conditioned on the output readout tokens.

Diffusion transformer architectures. RDT-1B [20] has been proposed as a large-scale diffusion transformer for robotics. In contrast to prior approaches, where the diffusion process is applied only at the action head, RDT-1B employs a Diffusion Transformer (DiT) [49] as its backbone, integrating the diffusion process directly into the transformer decoder to generate actions. In RDT-1B, language inputs are tokenized using the T5 encoder, while visual inputs are encoded using SigLIP. A diffusion model is then trained using a diffusion transformer with cross-attention, conditioned on both visual and textual tokens. To facilitate multimodal conditioning and avoid overfitting, Alternating Condition Injection is proposed, in which image and text tokens are alternately used as queries at each transformer layer.

Flow matching policy architectures. Recently, inspired by Transfusion [50], π_0 builds on PaliGemma [51] and introduces a custom action output module, the action expert, which enables a multimodal model to handle both discrete and continuous data. The action expert leverages flow-matching [52] to generate actions at rates up to 50Hz. It receives proprioceptive input from the robot and the readout token from the transformer, producing actions through a reverse diffusion process. Rather than generating tokens autoregressively, it outputs entire action chunks in parallel, enabling smooth and consistent real-time control.

Latent action learning from video. Another notable approach is LAPA [22], which leverages unlabeled video data for pre-training to learn latent actions for use in VLA models. This enables policies to effectively utilize human demonstrations, making them robust to changes in embodiment and well-suited for real-world deployment. The method applies patch embeddings, a spatial transformer, and a causal temporal transformer to images $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+H}$, then

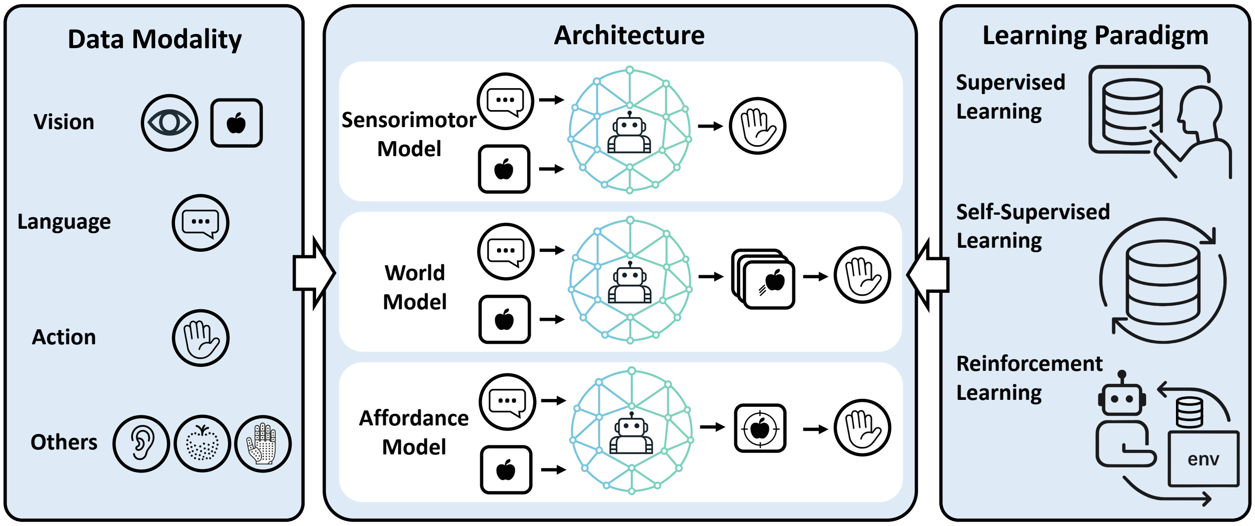


Figure 3: **Structure of Section 4 and Section 5.** The figure summarizes key components of VLA models. The center illustrates core architectural types, including sensorimotor models, world models, and affordance-based models. The left side depicts the input and output modalities—vision, language, action, and other auxiliary modalities. The right side presents training strategies, including supervised learning, self-supervised learning, and reinforcement learning, along with practical implementation considerations.

computes their difference. VQ-VAE [53] is applied to this difference, generating a discrete token z_t which, together with x_t, is used to reconstruct x_{t+H}. This entire network is trained jointly, forming a Latent Quantization Network. Building on LWM-Chat-1M (7B) [54], the vision and text encoders are kept frozen, and the resulting readout token is processed through an MLP trained to predict z_t. Finally, only the MLP component is replaced by a separate network trained to directly output robot control commands.

Hierarchical policy architectures. The most recent generation of VLAs adopts hierarchical policies to bridge high-level language understanding with low-level motor execution. RT-H [42] exemplifies this design by introducing a high-level controller that predicts intermediate "language motion" plans, followed by a low-level controller that refines these into concrete actions. The system can dynamically switch between generating symbolic actions and executing detailed control sequences, improving performance in long-horizon, multi-step tasks.

This design is extended in $\pi_{0.5}$ [55], which combines high-level action token generation (using FAST tokens) with a low-level controller trained via flow matching. Pre-training aligns symbolic actions with language, while post-training ensures smooth execution via continuous action decoding. GR00T N1 [24] integrates multiple elements: latent actions from LAPA, diffusion-based generation from RDT-1B, and flow-matching controllers from π_0, unified into a multi-stage policy that generalizes across robots and tasks. Hierarchical architectures now represent a state-of-the-art approach for scalable and adaptable VLA models, balancing the abstraction of language grounding with the precision of motor control.

4 Architectures and Building Blocks

Vision-Language-Action (VLA) models encompass a wide range of architectural designs, reflecting diverse strategies for integrating perception, instruction, and control. A widely adopted approach is the sensorimotor model, which jointly learns visual, linguistic, and action representations. These models take images and language as input and directly output actions, and can adopt either a flat or hierarchical structure with varying backbone architectures. While sensorimotor models form a foundational class of VLA systems, several alternative architectures have been proposed. World models predict the future evolution of sensory modalities, typically visual, conditioned on language input, and use these predictions to guide action generation. Affordance-based models are another variant that predict action-relevant visual affordances based on language, and then generate actions accordingly.

4.1 Sensorimotor Model

There are currently seven architectural variations of the sensorimotor models, as illustrated in Fig. 4.

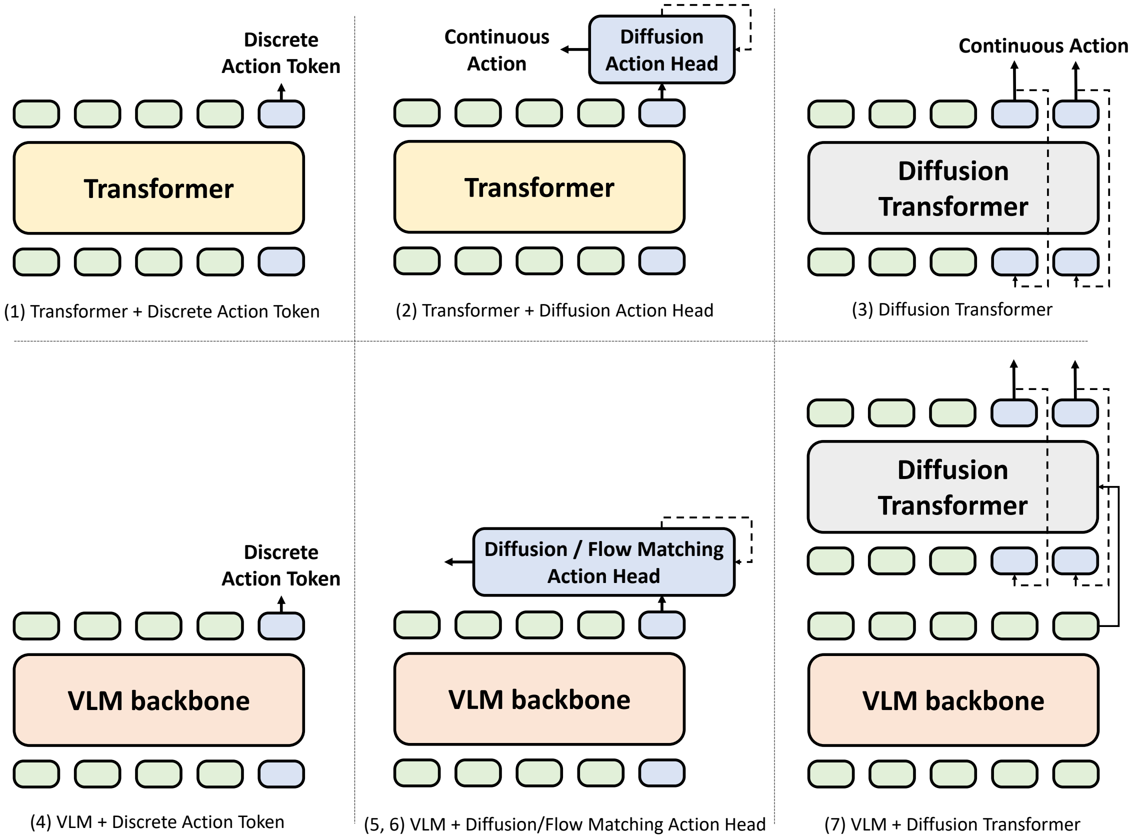

Discrete
Action Token
Transformer
(1) Transformer + Discrete Action Token
Continuous
Action
Diffusion
Action Head
Transformer
(2) Transformer + Diffusion Action Head
Continuous Action
Diffusion
Transformer
(3) Diffusion Transformer
Discrete
Action Token
VLM backbone
(4) VLM + Discrete Action Token
Diffusion / Flow Matching
Action Head
VLM backbone
(5, 6) VLM + Diffusion/Flow Matching Action Head
Diffusion
Transformer
VLM backbone
(7) VLM + Diffusion Transformer

Figure 4: **Architecture of sensorimotor models for VLA.** This figure categorizes seven representative architectures used in recent VLA research. (1) *Transformer + Discrete Action Token*: A standard transformer processes tokenized inputs to predict discrete actions. (2) *Transformer + Diffusion Action Head*: A diffusion model is appended to the transformer for generating smooth, continuous actions. (3) *Diffusion Transformer*: The diffusion process is integrated directly within the transformer architecture. (4) *VLM + Discrete Action Token*: Vision-language models (VLMs) replace transformers to leverage pre-trained knowledge while predicting discrete actions. (5) *VLM + Diffusion Action Head*: Combines VLMs with diffusion heads for continuous control. (6) *VLM + Flow Matching Action Head*: Substitutes diffusion with flow matching to enhance real-time control. (7) *VLM + Diffusion Transformer*: Employs a VLM as a backbone and a diffusion transformer as a low-level policy for end-to-end continuous action generation.

(1) Transformer + Discrete Action Token. This architecture represents both images and language as tokens, which are fed into a transformer to predict the next action, typically in the form of discrete tokens (see Fig. 4 (1)). This category also includes models that use CLS tokens and generate continuous actions through an MLP. Representative examples include VIMA [56] and Gato [27], which tokenize multiple modalities using language tokenizers, vision transformers, MLPs, and other components, and output discretized actions such as binned values. VIMA employs an encoder-decoder transformer conditioned on diverse task modalities, whereas Gato uses a decoder-only transformer that autoregressively processes all tokens in a single sequence.

In contrast to VIMA and Gato, which generate action tokens autoregressively, RT-1 [16] adopts a different approach by compressing inputs using TokenLearner [37] and employing a decoder-only transformer to predict all action tokens non-autoregressively. In practice, 48 tokens are fed into the transformer, and the final 11 tokens are extracted as action outputs. This architecture has been adopted by several approaches, such as MOO [57], RT-Sketch [58], and RT-Trajectory [59]. It has also become a common design choice in other VLA models such as Robocat [60], RoboFlamingo [61], and many others [62–67], due to its simplicity and scalability.

(2) Transformer + Diffusion Action Head. This architecture builds upon the structure in (1) by incorporating a diffusion policy as the action head following the transformer. While discrete action tokens often lack real-time responsiveness and smoothness, these models achieve continuous and stable action outputs using diffusion models [68]. Representative examples include Octo [19] and NoMAD [69]. Octo processes image and language tokens as a single sequence through a transformer, then applies a diffusion action head conditioned on the readout token. In contrast, NoMAD replaces the language input with a goal image, compresses the transformer output via average pooling, and uses the resulting vector to condition the diffusion model. TinyVLA [70], RoboBERT [71], and VidBot [72] also adopt this architecture.

(3) Diffusion Transformer. The diffusion transformer model shown in Fig. 4 (3) integrates the transformer and diffusion action head, executing the diffusion process directly within the transformer. This enables the model to perform the diffusion process conditioned directly on image and language tokens. For example, RDT-1B [20], built on this architecture, generates a sequence of action tokens via cross-attention with a vision and language query, which are subsequently mapped to executable robot actions through an MLP. Similarly, Large Behavior Models (LBMs) also adopt the diffusion transformer architecture and emphasize the importance of large-scale and diverse pre-training. In addition, StructDiffusion, MDT, DexGraspVLA, UVA, FP3, PPL, PPI, and Dita [73–80] uses this architecture.

(4) VLM + Discrete Action Token. VLM + Discrete Action Token models, as illustrated in Fig. 4 (4), improve generalization by replacing the transformer in (1) with a Vision-Language Model (VLM) pre-trained on large-scale internet data. Leveraging a VLM allows these models to incorporate human commonsense knowledge and benefit from in-context learning capabilities. For example, RT-2 uses large-scale VLMs such as PaLM-E or PaLI-X as the backbone, which processes image and language tokens as input and outputs the next action as discrete tokens. Furthermore, LEO, GR-1, RT-H, RoboMamba, QUAR-VLA, OpenVLA, LLARA, ECoT, 3D-VLA, RoboUniView, and CoVLA [18, 42, 81–89] adopt this architecture.

(5) VLM + Diffusion Action Head. VLM + Diffusion Action Head models, as shown in Fig. 4 (5), build on (2) by replacing the transformer with a VLM. This architecture combines VLMs, which enable better generalization, with diffusion models that generate smooth, continuous robot action commands. For example, Diffusion-VLA, DexVLA, ChatVLA, ObjectVLA, GO-1 (AgiBot World Colosseo), PointVLA, MoLe-VLA, Fis-VLA, and CronusVLA [90–98] adopt this architecture. HybridVLA [99] further combines (4) and (5) to both autoregressively generate discrete tokens as well as use a diffusion action head to generate continuous actions within a single model.

(6) VLM + Flow Matching Action Head. VLM + Flow Matching Action Head models, as shown in Fig. 4 (6), replace the diffusion model in (5) with a flow matching action head [52], improving real-time responsiveness while maintaining smooth, continuous control. A representative example is π_0, based on PaliGemma [51], which achieves control rates of up to 50 Hz. Other examples include GraspVLA, OneTwoVLA, Hume, and SwitchVLA [100–103]. $\pi_{0.5}$ [23] integrates the architectures of (4) and (6), supporting both discrete tokens and flow matching within a unified framework.

(7) VLM + Diffusion Transformer. VLM + Diffusion Transformer models, shown in Fig. 4 (7), combine a VLM with a diffusion transformer described in (3). The VLM typically serves as a high-level policy (system 2), while the diffusion transformer acts as a low-level policy (system 1). The diffusion transformer may be implemented using either diffusion or flow matching. A representative model is GR00T N1 [24], which applies cross-attention from the diffusion transformer to VLM tokens and generates continuous actions via flow matching. This design is also used in CogACT, TrackVLA, SmolVLA, and MinD [104–107].

4.2 World Model

World models are capable of anticipating future observations or latent representations based on the current inputs. Their forward predictive capabilities have made them increasingly central to VLA systems, where they support planning, reasoning, and control. In this section, we group these approaches into three types, as illustrated in Fig. 5.

(1) Action generation in world models. In contrast to models that directly generate actions, world models generate future visual observations, such as images or video sequences, which are then used to guide action generation. For example, UniPi [108] employs a diffusion model inspired by Video U-Net [109] to generate video sequences from an initial observation image and task instruction. Then, an inverse dynamics model (IDM) translates the predicted image sequence into low-level actions. This combination of visual prediction and IDM-based control is a common design pattern in model-based VLAs. Similarly, DreamGen [110] and GeVRM [111] predict future visual representations for action generation. HiP [112] extends this idea by incorporating subtask decomposition with a LLM, enabling the execution of longer-horizon behaviors. Dreamitate [113] finetunes Stable Video Diffusion [114] to synthesize a video of human using a tool for manipulation tasks. Then, given the generated video, MegaPose [115] estimates the 6-DoF pose of the tool so that the robot can follow the estimated tool poses. In contrast to generating full video sequences, SuSIE [116] predicts abstract subgoal images by using InstructPix2Pix [117] to generate intermediate goal images from the initial observation and task instruction, which are then used to condition a diffusion policy. CoT-VLA employs a similar approach for chain-of-thought reasoning (see Section 4.1 for further details.). LUMOS [118] also generates a goal image, but does so using a world model that takes low-level action commands as input. In LUMOS, a policy is trained to imitate expert demonstrations by interacting with the learned world model.

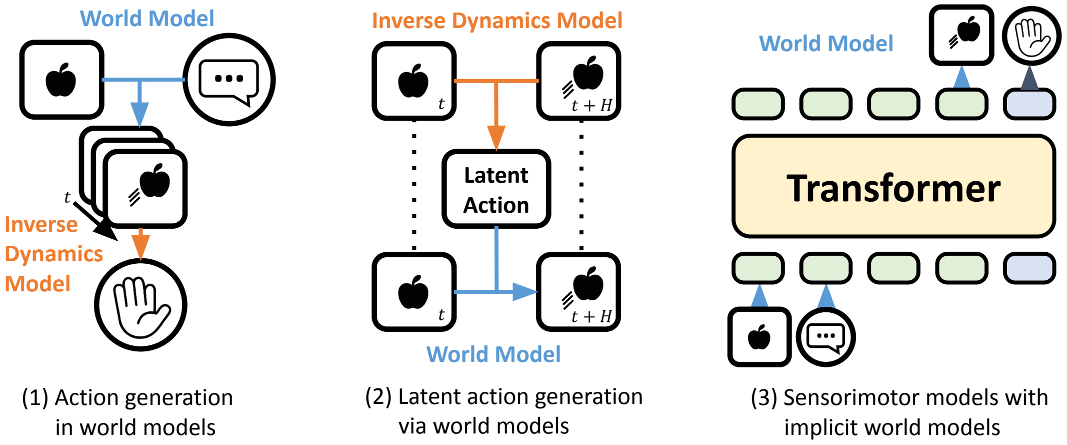
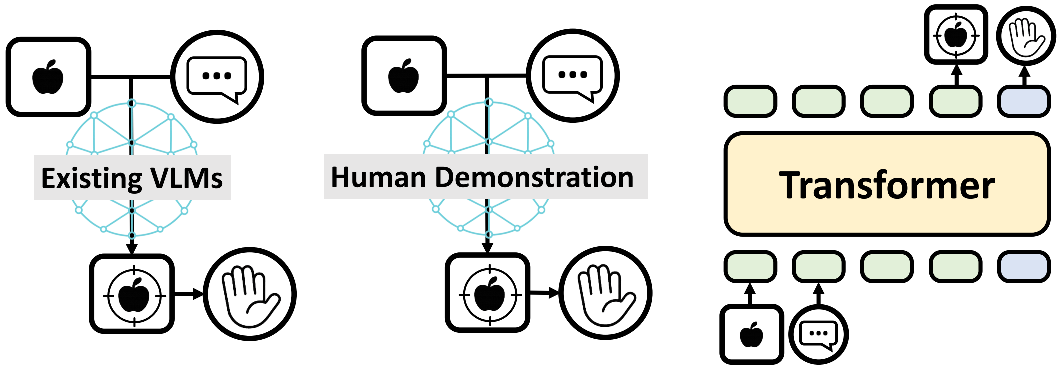

World Model
t
Inverse
Dynamics
Model
(1) Action generation in world models
Inverse Dynamics Model
t
t + H
Latent Action
t
t + H
World Model
(2) Latent action generation via world models
World Model
Transformer
(3) Sensorimotor models with implicit world models

Figure 5: **Design patterns for incorporating world models in VLA.** (1) Using world models in conjunction with inverse dynamics models to generate actions. (2) Leveraging world models to learn latent action representations, particularly from human videos; the resulting latent tokens are then used for VLA training to incorporate human video datasets. (3) Generating future observations in addition to actions, enabling predictive planning and multimodal reasoning.

In addition to video and image generation, many recent works leverage optical flow or feature point tracking. Because optical flow and feature tracking are agnostic to robot embodiment, they offer a more generalizable way to leverage human demonstrations. AVDC [119], similar to UniPi, generates video sequences and computes optical flow for each frame using GMFlow [120]. It then formulates the estimation of SE(3) rigid body transformations for target objects as an optimization problem. ATM [121] predicts future trajectories of arbitrary feature points (using CoTracker [122] during training), and trains a transformer that generates actions guided by these trajectories. Track2Act [123] predicts feature point trajectories between an initial and goal image, optimizes for 3D rigid body transformations, and learns a residual policy to refine the motion. LangToMo [124] predicts future optical flow from an initial image and task instruction, using RAFT [125] for optical flow supervision, and maps this prediction to robot actions. MinD [107] adopts an end-to-end approach that jointly learns video and action prediction. In particular, MinD combines a low-frequency video generator, which predicts future visual observations in a latent space from initial images and instructions, with DiffMatcher, which transforms these predictions into time-series features that the high-frequency action policy then uses to efficiently generate an action sequence. PPI [79] takes visual and language inputs to predict gripper poses and object displacements (Pointflow) at each keyframe. These are then used as intermediate conditions for action generation.

(2) Latent action generation via world models. This category of VLAs leverages world models to learn latent action representations from human demonstrations. For example, LAPA (Latent Action Pre-training from Videos) [22] (see Section 3 for details) jointly learns to predict action representations from tuples of current and future images, as well as to generate future frames conditioned on the current image and the latent action. This dual objective enables training on datasets without explicit action labels, such as human videos. Once latent actions are learned, a VLA policy is trained using these tokens. The action head is then replaced and fine-tuned to output robot actions. LAPA has been used for pre-training in GR00T N1 [24] and DreamGen [110]. Moreover, GO-1 [94] and Moto [126] employ a similar approach. UniVLA [127] augments the latent space of DINOv2 [46] with language inputs and uses a two-stage training process to disentangle task-independent and task-dependent latent action tokens. UniSkill [128] employs image editing based approach to extract latent actions from RGB-D images and uses them as conditions for a diffusion policy.

(3) Sensorimotor models with implicit world models. This category refers to VLAs that jointly output both actions and predictions of future observations to improve performance. GR-1 [82] integrates a pre-trained MAE-ViT encoder [129], CLIP text encoder [25], and a transformer, and is trained on the Ego4D dataset [130] to predict future observation images. It is then fine-tuned to jointly predict both actions and future frames from image, language, and proprioceptive inputs. By incorporating observation prediction, akin to a video prediction model, into a standard VLA framework, GR-1 demonstrates improved task success. GR-2 [131] builds on GR-1 by scaling up the training dataset and incorporating architectural improvements, including VQGAN-based image tokenization [132] and a conditional VAE [133] for action generation. GR-MG [134] generates intermediate goal images using an InstructPix2Pix-based



(1) Affordance prediction and action generation using VLMs

(2) Affordance extraction from human datasets

(3) Integration of sensorimotor models and affordance-based models

Figure 6: **Design patterns for incorporating affordance-based models in VLA.** (1) Predicting affordances and subsequently generating actions conditioned on the predicted affordances; (2) Extracting affordances from human demonstration videos and learning latent representations to guide action generation; (3) Integrating affordance prediction modules directly into the VLA architecture.

model [117] and embedding them within a GR-1-style framework. Furthermore, GR-3 [135] implements a hierarchical structure by integrating VLM (Qwen2.5-VL [136] and diffusion transformer with flow matching for action. 3D-VLA [87] extends this line of work by predicting RGB-D images with Stable Diffusion [137] and point clouds using Point-E [138]. Several other models incorporate full video prediction into sensorimotor models, including FLARE [139], UVA [76], WorldVLA [140], and ViSA-Flow [141].

4.3 Affordance-based Model

Affordances [142] refer to the action possibilities that an environment offers an agent, relative to its physical and perceptual capabilities. In robotics, this concept is often adapted to denote the actionable properties of objects or scenes, specifically, what actions are possible given the robot's embodiment and the spatial or functional cues present. VLAs based on affordance prediction can be currently categorized into three types, as illustrated in Fig. 6.

(1) Affordance prediction and action generation using VLMs. Pre-trained VLMs are often used to estimate affordances and generate corresponding actions. For example, VoxPoser [143] uses GPT-4 [2], OWL-ViT [144], and Segment Anything [145] to generate Affordance and Constraint Maps from language instructions, which are then used to guide action generation via Model Predictive Control (MPC). KAGI [146] employs GPT-4o [4] to infer a sequence of target keypoints from top-down and side-view images with overlaid grid lines, providing guidance for RL. LERF-TOGO [147] builds a 3D scene using a NeRF [148] trained on visual features extracted from CLIP and DINO [25,149] (LERF [150]). CLIP's text encoder is used to compute similarity between language instructions and visual features, and high-activation regions are converted into 3D point clouds, which are then processed by GraspNet [151] to rank grasp poses. Splat-MOVER [152] replaces NeRF with Gaussian Splatting [153] for faster scene construction and incorporates affordance heatmaps from the VRB model [154], improving both efficiency and performance.

(2) Affordance extraction from human datasets. This line of work focuses on extracting affordances from human motion videos, often without annotations, to enable scalable learning for robotic action generation. VRB [154] learns contact points and hand trajectories from demonstration videos in the EPIC-KITCHENS datasets [155,156]. In VRB, Hand-Object Detector (HOD) [157] is used to identify hand positions and contact states, then tracks subsequent hand movements on the image plane to automatically construct a training dataset. The extracted data are projected into 3D and used to generate robot actions. HRP [158] extracts hand, contact, and object affordance labels from the Ego4D dataset [130], trains a ViT model to predict these labels, and uses its latent representations for imitation learning. VidBot [159] extends 2D affordance representations to 3D, aiming to support zero-shot deployment on robots.

(3) Integration of sensorimotor models and affordance-based models. This approach incorporates affordance prediction into VLA. CLIPort [15] predicts affordances of objects and the environment from visual and language inputs, and generates actions based on these affordances. RoboPoint [160] builds a vision-language model that identifies affordance points, specific locations in an image where the robot should act, which are then projected into 3D to generate corresponding actions. RoboGround [161] predicts masks for the target object and placement area in pick-and-place tasks given image and language inputs; RT-Affordance [162] outputs key end-effector poses at critical moments; A_0 [163] predicts object contact point trajectories; and RoboBrain [164] identifies affordance regions as bounding boxes. Collectively, these models leverage affordance information as conditioning input for action generation. Chain-of-Affordance [165], inspired by Chain-of-Thought reasoning (see Section 4.1), predicts a sequence of affordances such as object positions, grasp points, and placement locations in an autoregressive manner, and then generates actions, leading to improved performance.

4.4 Data Modalities

VLAs process multiple modalities simultaneously, including vision, language, and action. This section summarizes how each modality is handled in state-of-the-art systems.

4.4.1 Vision

The most common approach for visual feature extraction in VLAs is to use ResNet [166] or Vision Transformer (ViT) [30]. These models are typically pre-trained on large-scale datasets such as ImageNet [167,168] or LAION [169, 170], although ResNet is often trained from scratch. Some methods apply ResNet directly to the image and convert the output into tokens using an MLP, while others first divide the image into patches before applying the encoder. Furthermore, ViT pre-trained with MAE [129] and EfficientNet [34] are also commonly used.

Vision-language models such as CLIP [25] and SigLIP [47] are also widely used. CLIP learns joint visual and textual representations via contrastive learning, while SigLIP improves upon it by removing the softmax constraint and reducing sensitivity to batch size. These models are often used alongside DINOv2 [46], a self-supervised vision model that learns image features without requiring paired text or contrastive objectives. While CLIP was initially the dominant choice, SigLIP and DINOv2 have emerged as the preferred models for visual feature extraction in VLAs. OpenCLIP [171] and EVA-CLIP [172] are also adopted in several prior works.

In addition, VQ-GAN [132] and VQ-VAE [53] are commonly used for discretizing images into token sequences. Unlike ViT or CLIP, which produce continuous embeddings, these models generate discrete tokens that are more naturally aligned with the input format of LLMs. The resulting visual tokens are often further processed to integrate with other modalities or to reduce token length. A well-known example is the Perceiver Resampler from Flamingo [173], which compresses visual information using a fixed-length set of learnable latent tokens via cross-attention. Building on this idea, Q-Former in BLIP-2 [3] combines cross-attention and self-attention to extract task-relevant information, while QT-Former [174] incorporates temporal structure into the process. TokenLearner [37] takes a different approach by performing spatial summarization to reduce token count. These compression and integration techniques are widely used in VLAs.

Several works in VLA adopt object-centric features, such as bounding box coordinates or cropped region embeddings, instead of relying solely on continuous feature maps. These features are typically extracted using object detection, segmentation, or tracking models, including Mask R-CNN [32], OWL-ViT [144], SAM [145], GroundingDINO [175], Detic [176], and Cutie [177].

4.4.2 Language

For language tokenization, VLAs typically inherit the tokenizer from their underlying LLM backbone, such as the T5 tokenizer [33] or LLaMA tokenizer [1]. When the base model is not a pre-trained LLM, tokenization is typically performed using subword algorithms such as Byte-Pair Encoding (BPE) or tools like SentencePiece [29], which implements BPE as well as other algorithms. For language encoding, VLAs employ various text encoders to embed natural language instructions into vector representations, including the Universal Sentence Encoder (USE) [36], CLIP Text Encoder [25], Sentence-BERT [178], and DistilBERT [179]. These language embeddings are frequently used to condition visual features via techniques such as FiLM conditioning. In architectures that use VLMs as backbones, visual information is directly integrated into the LLM component, with popular choices including LLaMA 2 [1], Vicuna [180], Gemma [181], Qwen2 [182], Phi-2 [183], SmolLM2 [184], GPT-NeoX [185], and Pythia [186].

4.4.3 Action

Action representation in end-to-end VLA models can be categorized into several primary approaches. This classification excludes specialized architectures such as affordance-based or world model-based methods.

Discretized action tokens obtained via binning. The most common approach to representing actions in VLAs is to discretize each dimension of the action space into bins (typically 256), with each bin ID treated as a discrete token. For example, RT-2 with PaLI-X [10, 39] directly outputs numeric tokens as actions; and RT-2 with PaLM-E [38] and OpenVLA [18] reserve the 256 least frequent tokens in the vocabulary for action representation. These models are typically trained using cross-entropy loss and adopt autoregressive decoding, similar to LLMs. Several models instead use non-autoregressive decoding, by inserting a readout token to enable parallel generation of all action tokens [187], or by treating the final few output tokens as discretized arm and base action (as in RT-1). A known drawback of standard binning is the increase in token length, which can limit control frequency. To mitigate this, FAST [55] applies the Discrete Cosine Transform (DCT) along the temporal axis, quantizes the frequency components, and compresses them using Byte-Pair Encoding (BPE). This significantly reduces token length and enables faster inference compared to conventional binning.

Decoding tokens into continuous actions. In this approach, tokens generated by a transformer are mapped to continuous actions via a multilayer perceptron (MLP), typically trained with an L2 or L1 loss. For binary outputs such as gripper open/close, binary cross-entropy is often used. OpenVLA-OFT [188] suggests that L1 loss may yield better performance. The MLP decoder can be replaced by alternative modules, such as an LSTM [189] to incorporate temporal context, or a Gaussian Mixture Model (GMM) to model stochasticity in the action space. Proprioceptive or force signals are often incorporated into the decoding module, such as an MLP or LSTM. Non-autoregressive variants commonly apply pooling operations (e.g., average or max pooling) to compress multiple tokens into a single action representation, as seen in RoboFlamingo [61]. OpenVLA-OFT [188] extends this by predicting multi-step action chunks, resulting in smoother and more temporally coherent trajectories.

Continuous action modeling via diffusion or flow matching. Diffusion models and flow matching have become prominent approaches for generating continuous actions in VLAs, as seen in Octo [19] and π_0 [21]. These models generate actions non-autoregressively, enabling smoother and more scalable control. Flow matching is particularly suitable for real-time applications, as it requires fewer inference steps than traditional diffusion. While some models implement diffusion as an external action head after the transformer, recent designs increasingly embed the process within the transformer itself, for example, in diffusion transformer architectures. Training and inference are commonly based on DDPM [68] and DDIM [190], with improved performance in stability and efficiency offered by methods such as TUDP [191], which ensure denoising consistency at every time step.

Learning latent action representations from web-scale data. This approach utilizes world modeling to obtain latent action representations when explicit actions are unavailable, such as in human demonstrations. By leveraging web-scale video data, this method enables training on significantly larger datasets and facilitates learning more generalizable VLAs. LAPA [22], Moto [126], UniVLA [127], and UniSkill [128] demonstrate this approach. For additional details, see Section 4.2.

Alternative action representation. SpatialVLA [192] statistically discretizes the action space and reduces the number of spatial tokens by allocating higher resolution to frequently occurring motions. ForceVLA [193] and ChatVLA [92] employ Mixture of Experts (MoE) architectures to dynamically switch action policies based on task phases. iManip [194] enables continual learning by incrementally adding learnable action prompts, preserving prior skills while acquiring new ones.

Cross-embodiment action representation. The challenge of embodiment diversity arises when handling robot-specific modalities such as actions and proprioception. Open X-Embodiment Project [17] was the first to tackle this embodiment challenge. Building upon the RT-1 [16] and RT-2 [10] architectures, this work standardized datasets across different robots using a unified format: single camera input, language instructions, and 7-DoF actions (position, orientation, and gripper open/close). This approach demonstrates a key insight that integrating data from robots with diverse embodiments leads to significantly improved VLA model performance compared to training on a single embodiment. Moreover, another prior work [195] has proposed to normalize and align actions and observations from heterogeneous embodiments into a shared first-person perspective, thereby enabling unified control of various robots using only observations and goal images [195]. However, such approaches struggle to uniformly handle robots with drastically different observations or control inputs, such as manipulators, mobile robots, and legged robots.

To address this limitation, CrossFormer [67] enables unified processing across diverse embodiments by first tokenizing heterogeneous sensor observations—such as vision, proprioception, and task specifications—using modality-specific tokenizers. All tokens are then assembled into a unified token sequence, with missing modalities masked as needed. This sequence is processed by a shared decoder-only transformer, which uses readout tokens to extract task-relevant representations. These are subsequently passed to embodiment-specific action heads (e.g., single-arm, bimanual, navigation, or quadruped) to generate actions tailored to each robot type. UniAct [196] proposes a Universal Action Space (UAS) implemented as a discrete codebook shared across embodiments. A transformer predicts discrete action tokens from this codebook, which are then converted into continuous actions by embodiment-specific decoders. By explicitly defining a shared atomic action space, UniAct facilitates knowledge transfer and promotes reusability across diverse robot embodiments. Furthermore, UniSkill [128] incorporates human demonstration knowledge by extracting latent skill representations from unlabeled human video data, in addition to robot data, similar to LAPA [22], enabling more generalizable VLA models. Additionally, embodiment-agnostic frameworks such as LangToMo [124] and ATM [121] achieve cross-embodiment learning by leveraging intermediate representations, such as optical flow and feature point trajectories, thereby bypassing the need for direct action space alignment.

4.4.4 Miscellaneous Modalities

In addition to vision, language, and action, modern VLA models increasingly incorporate additional modalities to enhance perception and interaction capabilities. In this section, we describe three additional sensing modalities relevant to VLA systems: audio, tactile sensing, and 3D spatial information.

Audio. Several prior works such as Unified-IO 2 [62], SOLAMI [197], FuSe [198], VLAS [199], and MultiGen [200] leverage audio information as input. Audio encoders typically take spectrograms or mel-spectrogram images as input, which are then converted into audio tokens using models like ResNet or ViT-VQGAN. RVQ-VAE-based SpeechTokenizer [201], Audio Spectrogram Transformer (AST) [202], or the Whisper encoder [203] are also frequently used as pre-trained models. These encoders enable the system to leverage rich audio information that may not be readily transcribed into text for robotic decision-making. SoundStorm [204] or the decoder of VQGAN are often employed for decoding. A common and straightforward approach, as employed in RoboNurse-VLA [205], is to convert audio into text using standard automatic speech recognition (ASR) systems.

Tactile sensors. FuSe [198], TLA [206], VTLA [207], and Tactile-VLA [208] incorporate tactile information as part of inputs. Tactile sensors such as DIGIT [209] and GelStereo 2.0 [210], which produce image-based outputs, are commonly used. These tactile images are either encoded using a Vision Transformer (ViT) or tokenized via a pre-trained Touch-Vision-Language (TVL) model [211]. This enables the integration of visual and tactile information for learning fine-grained manipulation skills in contact-rich tasks, such as peg insertion. Although not tactile sensors in the strict sense, ForceVLA [193] incorporates general 6-axis force-torque sensors. In particular, a force-aware Mixture-of-Experts fusion module integrates force tokens derived from 6-axis force–torque sensor data with visual-language features extracted by a pre-trained VLM, and generates actions through an action head.

3D information. Incorporating 3D information enables robots to more accurately perceive their environment and plan actions accordingly. In 3D perception, we specifically introduce (a) depth images, (b) multi-view images, (c) voxel representations, and (d) point clouds below.

(a) Depth images. A common strategy for incorporating depth information involves tokenizing depth images using standard visual backbones, such as Vision Transformers (ViTs) or ResNets, similar to the processing of RGB images. In scenarios where direct depth sensing is not available, monocular depth estimation models such as Depth Anything [212] and ZoeDepth [213] are frequently utilized to predict depth from RGB inputs. SpatialVLA [192] is a representative method that utilizes depth images by introducing Ego3D Position Encoding. In this framework, depth maps are first estimated from RGB inputs using ZoeDepth, and the corresponding 3D coordinates for each pixel are computed via the camera's intrinsic parameters. The 3D coordinates are first processed using sinusoidal positional encoding and an MLP, and the resulting features are added to the 2D visual features extracted by SigLIP [47]. This combined representation is used as the Ego3D positional encoding and provided as input to the LLM. Additionally, HAMSTER [214], RationalVLA [215], and OpenHelix [216] incorporate a 3D Diffuser Actor [217], a diffusion-based action head that operates in 3D space and processes RGB-D inputs to generate actions.

(b) Multi-view images. Several works attempt to extract 3D information from multi-view images. For example, GO-1 [94] simply takes as input multi-view RGB-D images, encouraging implicit understanding of 3D structure. 3D-VLA [87] extends Q-Former (described in Section 4.4.1) to handle RGB-D and multi-view inputs. Evo-0 [218] employs Visual Geometry Grounded Transformer (VGGT) [219] to extract implicit 3D geometric information from multi-view RGB images. RoboUniView [88] and RoboMM [220] utilize UVFormer, a pre-trained model that takes multi-view RGB-D images and corresponding camera parameters as input and outputs a 3D occupancy grid. The encoder's output features are then used as tokens for downstream processing. Furthermore, SAM2Act [221] and HAMSTER [214] use Robotic View Transformer-2 (RVT-2) [222] to reproject point cloud or depth information into a virtual view (often using orthographic projection to generate three images), and each image is tokenized by ViT. Similar approaches are also used in OG-VLA [223] and BridgeVLA [224]. Overall, two main approaches have emerged: integrating information from multiple viewpoints, and projecting 3D data into orthographic images to facilitate easier processing.

(c) Voxel representations. Voxel-based representations are another widely adopted approach for encoding 3D information. OccLLaMA [225] and OpenDriveVLA [226] convert 3D occupancy grids into 2D Bird's Eye View (BEV) feature maps, which are then tokenized using VQ-VAE. Several approaches operate directly on three-dimensional voxel grids, such as iManip [194], which extracts features using a 3D U-Net [227], and VidBot [72], which first converts voxel grids into Truncated Signed Distance Fields (TSDFs) and then processes them using a 3D U-Net. Because voxel representations resemble image structures and are compatible with convolutional processing, they have been widely adopted across various studies.

(d) Point clouds. A common approach involves tokenizing point clouds using pre-trained point-based transformers such as PointNet [228], PointNet++ [229], PointNext [230], and Uni3D ViT [231]. These backbones are widely adopted in models such as SOFAR [232], LEO [81], PPI [79], LMM-3DP [233], GeneralFlow [234], FP3 [77], and DexTOG [235]. In contrast, some methods opt for task-specific training: StructDiffusion [73] uses the Point Cloud Transformer (PCT) [236], and PointVLA [95] employs PointCNN [237], with both models trained from scratch for their respective tasks. Additionally, although less common, LERF-TOGO [147] and Splat-MOVER [152] integrate point clouds reconstructed using Neural Radiance Fields (NeRF) or Gaussian Splatting with semantic features extracted from CLIP [25]. These enriched representations are then used in conjunction with GraspNet [151] to generate grasping plans.

Beyond the primary modalities discussed above, several VLA models have been proposed to incorporate additional forms of information. ARM4R [238], for example, integrates 3D tracking data to enhance motion understanding. SOLAMI [197] introduces a Motion Tokenizer that applies VQ-VAE to discretize the joint angles of SMPL-X [239] on a per-body-part basis, following the approach introduced in motionGPT [240]. Additionally, PPL [78] and LangToMo [124] incorporate motion dynamics by using RAFT [125] to estimate optical flow from pairs of images, enabling fine-grained temporal reasoning.

4.5 Emerging Techniques

Recent advances in VLA research highlight two emerging directions: *hierarchical architectures* and *Chain-of-Thought (CoT) reasoning*. Both approaches introduce structured intermediate representations between language instructions and low-level actions, enabling more robust planning, decomposition, and reasoning. Further details of these approaches are provided below.

Hierarchical architectures. The most foundational approach is Atomic Skill [241] and LMM-3DP [233], which use existing VLMs as high-level policies to decompose task instructions into subtasks. These subtask descriptions are then passed to a VLA acting as the low-level policy. Since the low-level policy receives cleaner and more concise language inputs, it can execute actions more reliably than when processing complex, unstructured instructions directly. On the other hand, Hi Robot [242] trains a custom high-level policy instead of relying on existing VLMs. NAVILA [243] and HumanoidVLA [244] employ low-level policies trained using reinforcement learning (RL) to achieve fine-grained motor control. RT-H [42] and LoHoVLA [245] take a more integrated approach by jointly training both high-level and low-level policies within a single network. By switching the input prompt, these models can flexibly alternate between decomposing a task instruction into subtasks and converting a subtask into a corresponding action. This approach has been further extended to $\pi_{0.5}$ [23], which unifies subtask decomposition, discrete action token generation, and continuous action generation within the same network. The integration of task decomposition with VLA models is emerging as a promising approach for enabling more flexible and scalable robot behavior. Additionally, FiS-VLA [97], OpenHelix [216], and DP-VLA [246] propose connecting high-level and low-level policies through latent spaces, without explicitly defining intermediate representations as subtasks. Tri-VLA [247] integrates a pre-trained vision-language model for scene understanding with Stable Video Diffusion, which produces visual representations capturing both static observations and future dynamics. These representations are then used as input to a diffusion transformer, which generates actions via cross-attention.

Chain-of-Thought (CoT) reasoning. Chain-of-Thought (CoT) reasoning, while conceptually similar to hierarchical approaches, introduces a distinct mechanism that has been integrated into VLA models such as ECoT [86] and CoT-VLA [187]. ECoT addresses a key limitation of typical VLAs, which is their lack of intermediate reasoning, by introducing a step-by-step process between observations, instructions, and action generation, thereby enhancing planning and inference capabilities. In particular, ECoT achieves this by autoregressively predicting intermediate representations, such as task descriptions, subtasks, and object positions, before generating the final action sequence. On the other hand, CoT-VLA [187] generates subgoal images, thereby improving success rates on more visually grounded tasks. ECoT-Lite [248] reduces inference latency caused by reasoning by selectively dropping certain reasoning components during training. Fast ECoT [249] takes this further by reusing intermediate reasoning outputs and parallelizing reasoning and action generation, resulting in faster action execution.

5 Training Strategy and Implementation

We categorize the training approaches of Vision-Language-Action (VLA) models into supervised learning, self-supervised learning, and reinforcement learning. Below, we summarize the core characteristics and representative methods of each approach.

5.1 Supervised Learning

Most VLA models are trained using supervised learning on datasets consisting of pairs of images, language, and actions. Since many VLAs are built on LLMs, training is often formulated as a next-token prediction task. The choice of action loss function depends on the architecture of the action head, such as MLPs, diffusion models, or flow matching networks, ensuring appropriate supervision for each model type.

VLA training generally consists of two stages: pre-training and post-training. In many cases, a LLM or VLM pre-trained on web-scale data is first used as the initial backbone for training. While some models are trained from scratch, it is more common to initialize training with a pre-trained VLM that has already acquired commonsense knowledge, in order to enhance generalization. Pre-training is typically performed using datasets such as human demonstrations, heterogeneous robot demonstrations, or VQA datasets related to robotic planning. Similar to LLMs, data scale plays a crucial role in VLA pre-training. Leveraging large and diverse datasets enables the development of VLA models with stronger generalization across tasks and embodiments. In the pre-training stage, the pre-trained VLM is typically fully fine-tuned to adapt to robotics-related domains. For further details about pre-training, see Section 5.4.1.

After pre-training, post-training is performed using task- or robot-specific datasets. In this stage, data quality tends to be more important than quantity, and the datasets are often smaller to those used in pre-training. Finetuning strategies differ across implementations. In some cases, the entire model undergoes full finetuning, whereas in others, adaptation is limited to the action head.

Moreover, in-context learning, a technique originally developed for LLMs, has also been adapted for use in VLA systems. Rather than explicitly fine-tuning on demonstration data, in-context VLA models condition on a small number of human teleoperation trajectories at test time to infer appropriate actions. For instance, ICRT [250] introduces a framework in which 1–3 teleoperated demonstrations are provided as prompts, enabling the model to generate corresponding robot actions in a zero-shot manner.

5.2 Self-Supervised Learning

Self-supervised learning is occasionally incorporated into the training of Vision-Language-Action (VLA) models, serving three primary purposes.

Modality alignment focuses on learning temporal and task-level consistency across modalities in VLA models. For instance, TRA [251] uses contrastive learning to align representations of current and future states within a shared latent space, achieving temporal alignment. Similarly, task alignment is achieved by aligning language instruction embeddings with those of goal images through contrastive objectives.

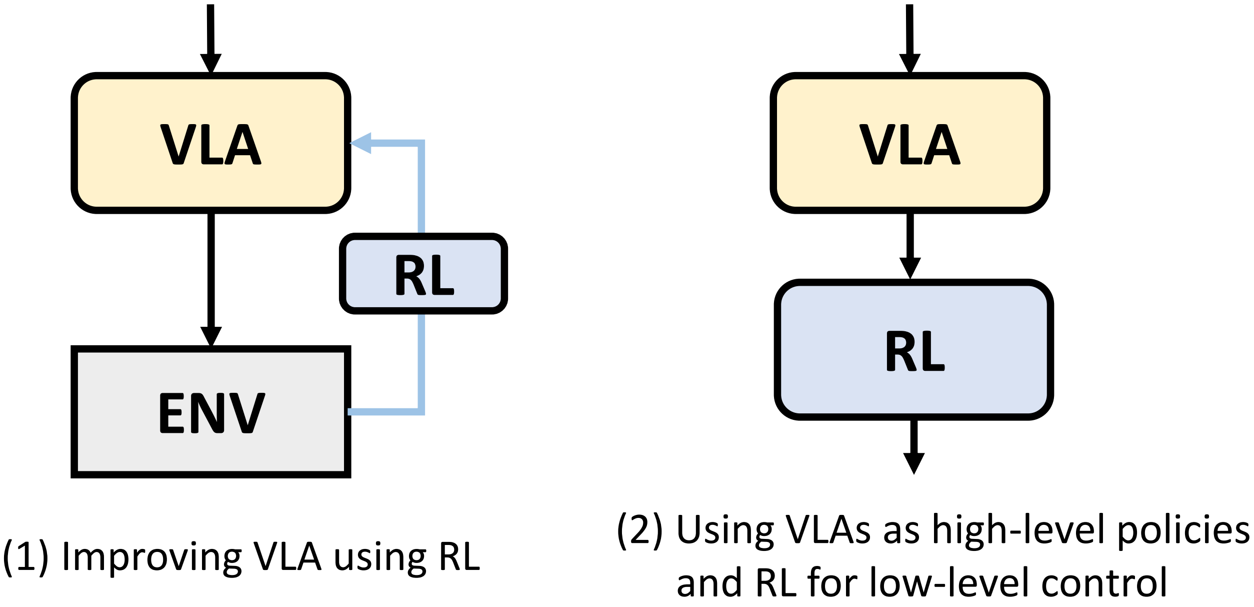
Visual representation learning aims to extract visual features from images or videos using techniques such as masked autoencoding (e.g., MAE [129]), contrastive learning (e.g., CLIP [25]), and self-distillation (e.g., DINOv2 [46]). These pre-trained models are widely adopted in VLAs as foundational visual encoders.

Latent action representation learning leverages self-supervised techniques to learn action embeddings, as discussed in Section 4.2 and Section 4.4.3. By extracting a latent action from the initial and goal images, and reconstructing the goal image using the initial image and the extracted latent action, the model learns meaningful action representations without requiring explicit labels. This approach is highly scalable and well-suited for large, unannotated datasets.

5.3 Reinforcement Learning

While VLA is trained via imitation learning in general, imitation learning alone faces challenges such as the inability to handle novel behaviors and the requirement for sufficiently large and high-quality expert demonstrations. To address these issues, several prior arts have explored finetuning VLA or training low-level policies using reinforcement learning (RL), such as PPO [252] and SAC [253]. These approaches can be broadly categorized into the following two types, as shown in Fig. 7.

(1) Improving VLA using RL. Recent work leverages RL to improve the robustness, adaptability, and real-world performance of VLA models. Several approaches fine-tune VLAs using RL with task success or failure as the reward



(1) Improving VLA using RL

(2) Using VLAs as high-level policies and RL for low-level control

Figure 7: **Approaches to integrating RL with VLA models.** (1) RL is used to fine-tune VLA models to enhance their performance. (2) VLA models serve as high-level policies, while RL policies handle low-level control.

signal. iRe-VLA [254] achieves high performance by repeatedly combining supervised fine-tuning (SFT) on expert data, online RL on the action head using success and failure rewards, and subsequent SFT using both expert data and successful trajectories collected during online learning. ConRFT [255] applies imitation learning on a small set of demonstrations, performs offline RL to learn a Q-function, and subsequently fine-tunes the policy online through human interventions. This approach is inspired by prior frameworks such as SERL [256] and HIL-SERL [257], which are reset-free [258, 259], off-policy RL methods [260] designed for real-world robot learning. VLA-RL [261] introduces the Robotic Process Reward Model (RPRM), which replaces sparse binary rewards with dense pseudo-rewards derived from gripper actions and task progress, enabling more stable PPO-based training. RLDG [262] fine-tunes large VLA models such as OpenVLA [18] and Octo [19] using successful trajectories collected via HIL-SERL, allowing integration of multiple expert policies into a unified VLA. MoRE introduces a Mixture of Experts (MoE) structure into the VLA, enabling token-wise expert selection and refinement via RL. RLRC [263] compresses OpenVLA by pruning up to 90% of its parameters, recovers performance via SFT, and then applies RL for final fine-tuning using task-level feedback. These studies demonstrate that RL, especially when combined with expert demonstrations or human interventions, can significantly improve the flexibility and reliability of VLA models in real-world settings. More recently, to address the potential instability associated with backpropagation through diffusion chains, DSRL [264] proposes applying RL in the latent noise space of the diffusion policy. This approach avoids updating the parameters of the underlying VLA model during RL fine-tuning. Instead, it learns a distribution over the latent noise, allowing the model to sample informative initial noise vectors rather than from a standard Gaussian. Notably, DSRL demonstrates that the success rate of π_0 can be improved from approximately 20% to nearly 100% using only 10K samples.

(2) Using VLAs as high-level policies and RL for low-level control. This class of approaches delegates high-level decision-making to the VLA, while low-level control is handled by policies trained with RL. Humanoid-VLA [244] uses a VLA to generate high-level commands, which are executed by a whole-body controller trained via RL for humanoid robots. NaVILA [243] adopts a similar strategy, applying RL to convert velocity commands from the VLA into torque control for a legged robot. A more advanced system, SLIM [265], targets a mobile manipulator comprising a quadruped base and robotic arm. It first trains a teacher policy using RL with privileged inputs, such as footstep plans, object placements, and subtask identifiers, to generate base and arm trajectories. This policy is then distilled into a student VLA via imitation learning, enabling end-to-end mapping from images and language to actions. RPD [266] takes a complementary approach, using a pre-trained VLA to guide exploration during RL. Here, the VLA acts as a teacher, shaping the learning process rather than serving as a high-level controller.

In addition, LUMOS performs imitation learning in the latent space of a world model by employing reinforcement learning guided by an intrinsic reward that quantifies the deviation from expert trajectories within the latent space. DexTOG [235] generates a diverse set of grasp poses using a diffusion model and employs reinforcement learning to evaluate whether each candidate pose leads to task success. Through iterative fine-tuning with successful trajectories, the diffusion model learns object-specific grasp poses that are well-suited for subsequent tasks.

Despite the growing number of VLA methods incorporating RL, most prior work remains limited to simulation or simplified real-world setups, due to sample inefficiency, unsafe exploration, and computational inefficiency.

5.4 Training Stages

Training Vision-Language-Action (VLA) models typically involves multiple stages, each targeting a specific aspect of learning. The pre-training aims to acquire general capabilities and promote transferability across diverse robotic embodiments. When a pre-trained Vision-Language Model (VLM) is used as the backbone of a VLA model, it must be adapted to the robotics domain to effectively ground language and visual understanding in action. This is followed by a post-training, in which the model is further refined using high-quality robot demonstration data to improve performance on specific downstream tasks. This section provides a stage-wise overview of representative training strategies, highlighting common data sources, model backbones, and adaptation techniques used in recent VLA systems.

5.4.1 Pre-training

Pre-training plays a pivotal role in shaping the generalization ability and semantic grounding of VLA models. This subsection outlines key strategies and design choices in recent pre-training pipelines, highlighting how large-scale multimodal data, powerful VLM backbones, and training stabilization techniques contribute to effective policy initialization.

Data scale and source. The scale and heterogeneity of training data significantly impact the generalization ability of VLA models across diverse scenes, objects, and tasks. Recent models increasingly leverage large-scale datasets that combine robot demonstrations, web-scale vision-language corpora, and structured annotations to improve semantic understanding and visuomotor grounding.

π_0 [21] is trained on millions of real-world trajectories collected across varied embodiments and tasks. Its successor, $\pi_{0.5}$ [23], extends this approach by incorporating not only robotic data but also large-scale vision-language datasets commonly used for object detection and visual reasoning (e.g., COCO [267], VQA [268]). The model is trained with auxiliary cross-entropy losses for multiple tasks, including bounding box prediction, image captioning, subtask language generation, and discrete action prediction.

Similarly, Gr00T N1 [24] incorporates an auxiliary bounding box loss to improve spatial localization and affordance detection. These bounding box labels are obtained using OWL-ViT [144], allowing the model to learn from weakly supervised visual data. Gr00T N1 further leverages egocentric human videos, from which latent action representations are extracted to supervise the VLA model. Additionally, it introduces diverse synthetic trajectories generated in simulation, which are transformed into realistic visual observations using the COSMOS world model [269], enhancing the model's capacity to learn long-horizon, multi-stage behaviors.

These approaches demonstrate a growing trend toward enriching VLA training data not only in scale but also in structure and modality. By jointly training on action, grounding, and reasoning tasks, modern VLAs acquire richer representations that support robust policy learning and generalization.

VLM backbones. A common practice in recent VLA models is to leverage vision-language models (VLMs) that have been pre-trained on large-scale web data. This strategy enables models to inherit broad visual and linguistic priors, including common sense knowledge, semantic grounding, and reasoning capabilities. By decoupling low-level perceptual grounding from action policy learning, pre-trained VLMs provide a flexible foundation that can be adapted to various robotic tasks with limited additional supervision. We now introduce a selection of representative VLM backbones that have been employed in VLA models.

- **PaLM-E** [38], developed by Google, has been used—along with PaLI-X [39]—as the backbone for RT-2 and its successor VLA models.
- **PaliGemma** [51] combines Gemma [181] with SigLIP [47], and is used in π_0 [21] and $\pi_{0.5}$ [23] developed by Physical Intelligence.
- **PrismaticVLM** [45] is based on LLaMA 2 [1] and combines it with DINOv2 [46] and SigLIP [47]. It is widely used in current VLA models, including OpenVLA [18] and CogACT [104].
- **Qwen2.5-VL** [136], developed by Alibaba, combines Qwen2.5 LLM [270] with a ViT-based vision encoder. It is used in a variety of VLA models such as NORA [271], Interleave-VLA [272], and CombatVLA [273].

- **LLaVA** [274] integrates the LLaMA-based LLM Vicuna [180] with the vision encoder from CLIP [25] via an MLP. It has been widely adopted in models such as OpenHelix [216], OE-VLA [275], and RationalVLA [215].
- **Gemini 2.0** [276], developed by Google, includes variants such as Gemini Robotics-ER for robotic question answering and Gemini Robotics, which extends its capabilities to VLA applications [277].
- **Fuyu-8B** [278]: QUAR-VLA [279] and MoRE [280],
- **OpenFlamingo** [61]: RoboFlamingo [61], DeeR-VLA [281], and RoboMM [220],
- **BLIP-2** [3]: 3D-VLA [87],
- **LLaMA3.2** [282]: FOREWARN [283],
- **AnyGPT** [284]: SOLAMI [197],
- **Phi** [183]: TraceVLA [285], UP-VLA [286], and HybridVLA [99],
- **Molmo** [287]: UAV-VLA [288],
- **VILA** [289]: NaVILA [243] and HAMSTER [214],
- **InternVL** [290] GO-1 [94],
- **Eagle-2** [291]: GR00T N1 [24],
- **Chameleon** [292]: WorldVLA [140].

This demonstrates the extensive diversity in VLM backbones currently employed across the VLA landscape.

Gradient insulation. An emerging trend in training VLA models involves preventing gradient flow from the action head into the vision-language backbone [293]. Allowing gradients from a randomly initialized action head to propagate can compromise pre-trained representations, resulting in unstable and inefficient training. Prior work demonstrates that this form of gradient insulation significantly improves both training stability and efficiency [293]. GR00T N1.5 [24] also freezes the VLA model entirely, likely for similar reasons. Similarly, RevLA [294] also addresses catastrophic forgetting by gradually reversing the backbone model weights, inspired by model merging.

Stability and efficiency heuristics. Re-Mix [295] adjusts the sampling weights of individual datasets based on excess loss, which quantifies the remaining potential for policy improvement within each domain.

5.4.2 Post-training

In contrast to pre-training, which relies on large-scale and diverse datasets, post-training requires high-quality, robot- and task- specific data. As full fine-tuning typically demands substantial computational resources, an alternative strategy is to fine-tune only the action head while keeping the backbone weights frozen. Another approach is to use Low-Rank Adaptation (LoRA) [296], which enables computationally efficient fine-tuning with minimal performance degradation.

In addition, BitVLA [297] introduces a distillation-based approach to quantize the vision encoder, aiming to enable memory-efficient training. Specifically, the vision encoder is compressed to 1.58 bits by distilling a full-precision encoder into a quantized student model. This strategy achieves substantial memory savings with minimal performance degradation, thereby facilitating efficient deployment on resource-constrained systems.

Freezing backbone vs. full fine-tuning. When adapting pre-trained VLMs for robotic tasks, a critical design choice is whether to freeze the vision-language backbone or perform full fine-tuning. This decision involves fundamental trade-offs across multiple dimensions.

(a) Computational efficiency: Freezing the backbone requires significantly less GPU memory and training time as gradients only need to be computed for the action head, enabling training on consumer-grade GPUs. In contrast, full fine-tuning demands substantial computational resources, often requiring large GPU clusters and extended training periods, which limits accessibility for many researchers.

(b) Domain adaptation: Full fine-tuning excels by enabling end-to-end optimization that jointly learns perception and control, allowing the model to adjust to robot-specific visual patterns and domain-specific knowledge. Frozen backbones, however, cannot adapt to these domain shifts, potentially creating a gap between pre-trained representations and robotic perception requirements.

(c) Performance-resource trade-off: Full fine-tuning of VLA models often yields the highest task-specific performance when sufficient data and compute are available, but it incurs substantial computational cost. To mitigate this, parameter-efficient adaptation methods such as Low-Rank Adaptation (LoRA) [296] offer a compelling alternative.

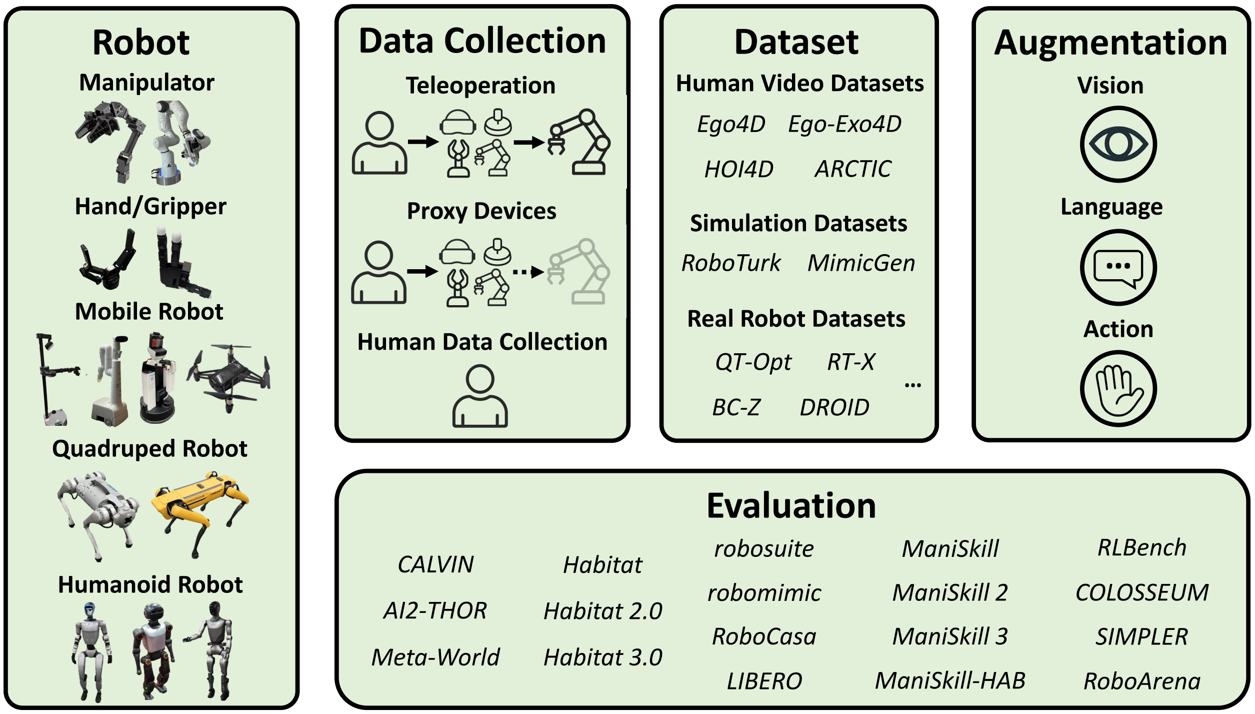


Figure 8: **Structure of Section 6 and Section 7:** robots used VLA research — including manipulator, hand/gripper, mobile robot, quadruped robot, and humanoid robot; data collection methods — including teleoperation, proxy devices, and human data collection; publicly available dataset — including human egocentric data, simulation data, and real-world robot data; augmentation for vision, language, and action; various evaluation benchmarks.

For instance, OpenVLA [18] demonstrates that LoRA can achieve competitive performance while significantly reducing memory and compute requirements, enabling training on consumer-grade GPUs rather than large-scale clusters. Recent work has also explored intermediate strategies, such as staged unfreezing or selective fine-tuning of specific layers, to strike a balance between adaptation capability and efficiency.

(d) Knowledge preservation: Frozen backbones maintain the rich visual and linguistic representations learned from web-scale data, preventing catastrophic forgetting of general vision-language capabilities. Full fine-tuning, while allowing the model to specialize for robotic visual features and action-grounded language, risks degrading these pretrained representations, potentially losing valuable general knowledge that could benefit zero-shot generalization.

5.5 Inference

To address latency during real-world execution, Real-Time Chunking (RTC) [298] introduces an asynchronous action generation strategy. RTC mitigates delays by fixing previously executed actions while generating subsequent actions in the sequence. This method uses soft masking to maintain temporal consistency with past trajectories while enabling dynamic replanning based on updated sensory inputs.

Furthermore, DeeR-VLA [299] is trained to enable action prediction at each layer of the transformer. If the difference between actions predicted from two consecutive layers is small, the remaining layers are skipped to accelerate inference. VLA-Cache [300] improves inference speed by identifying static tokens and reusing previously computed features from earlier steps.

6 Datasets

6.1 Data Collection for VLA

Training VLA models requires access to large volumes of high-quality data. This section outlines the primary data collection strategies employed in VLA research. Note that data collection via simulation is discussed in Section 6.2; here we focus on methods based on real devices.

Teleoperation. In this approach, demonstrations are recorded in real time while a human operator directly controls the robot, enabling the collection of high-quality trajectories. This method forms the basis of many VLA datasets. For example, ALOHA [301] employs a unilateral teleoperation setup consisting of a dual-arm WidowX 250 as the leader and a dual-arm ViperX-300 as the follower. The follower robot mimics the leader's motions, allowing precise manipulation data to be captured. Mobile ALOHA [302] extends this framework by mounting the system on a mobile base, enabling the collection of mobile manipulation demonstrations. The ALOHA framework has evolved through multiple iterations. ALOHA 2 introduces refined hardware components, such as upgraded grippers and gravity compensation mechanisms, along with open-source hardware and simulation environments [303]. Building on this upgraded platform, ALOHA Unleashed investigates large-scale imitation learning [304]. Furthermore, Bi-ACT [305] introduces bilateral control to enable more responsive interaction between the leader and follower robots, while GELLO [306] adapts the system by employing a scaled-down follower robot with proportionally adjusted link lengths.

In contrast to leader-follower approaches, which require robots on both the leader and follower sides, many prior works have proposed methods to reduce both the burden on the human operator and the overall cost of the teleoperation system. For instance, AnyTeleop [307] estimates the position and orientation of the human hands from a single RGB camera using MediaPipe [308], and retargets this information to the robot via CuRobo [309] for teleoperation. ACE [310] combines precise wrist tracking using an exoskeleton device with hand pose estimation from Mediapipe to facilitate accurate teleoperation. Aiming for applications in humanoid robotics, Open-Television [311] utilizes hand and head pose estimation via the Apple Vision Pro to enable both teleoperation and active-vision-based manipulation. Bunny-VisionPro [312] also employs the Apple Vision Pro, with greater emphasis on haptic feedback and real-time system integration.

In addition to these approaches, data collection can also be performed through more direct control methods such as 3D mice or game controllers. While these alternatives simplify the setup and eliminate the need for wearable or vision-based pose estimation systems, they may offer lower fidelity in replicating natural human motions.

Data collection using proxy devices. Controlling a physical robot directly poses significant challenges for scaling data collection. By decoupling human motion from physical robot control, recent approaches enable more intuitive, flexible, and scalable data collection through the use of proxy devices. For example, UMI [313] is a handheld gripper equipped with a GoPro camera, whose 6-DoF trajectory is estimated using visual SLAM. The collected data can be used to train a policy, and by later mounting UMI as the robot's end-effector, the robot can reproduce the demonstrated motions without being physically involved during data collection. Recently, LBM [314] leverages UMI to collect 32 hours of demonstrations. DexUMI [315] extends this concept to dexterous manipulation by replacing the simple gripper with a five-fingered robotic hand. The human demonstrator wears an exoskeleton glove equipped with the same cameras and tactile sensors as the target robot, allowing the recorded hand motions to be faithfully transferred.

Building on similar principles, Dobb-E [316] uses a rod-shaped device resembling the end-effector of Hello Stretch to capture human demonstrations. RUMs [317] further enhances this paradigm by increasing the diversity of collected tasks, incorporating failure detection mechanisms, and improving the network architecture. These improvements enable the robot to generalize to a wide range of tasks through pre-training alone. DexCap [318] is the device for data collection by mounting Realsense T265 cameras on Rokoko EMF gloves for both hands, along with additional Realsense T265 and L515 sensors on the chest, enabling SLAM-based 6-DOF wrist pose estimation and glove-based hand pose tracking. In contrast, DexWild [319] addresses the wiring complexity and SLAM calibration challenges of DexCap by using EMF gloves in combination with palm-facing cameras on both hands and ArUco marker tracking via external cameras.

Human data collection. This approach involves collecting data by recording natural human behavior without relying on proxy devices that mimic the robot's end effector. The simplest form of this method involves mounting a GoPro camera or microphone on the user's head to capture first-person visual and auditory data, often supplemented with inertial measurement unit (IMU) or gaze information. This technique has been widely adopted in large-scale egocentric datasets such as Ego4D and EPIC-KITCHENS [130,155,156]. Recent advances in wearable sensing technologies have enabled more naturalistic and scalable data collection using compact smart glasses such as Meta's Project Aria. These devices have facilitated the development of enriched datasets including Ego Exo4D, HOT3D, HD EPIC, and Aria Everyday Activities [320–323]. Leveraging these datasets, several prior works have trained robot policies directly from human demonstration data. For instance, EgoMimic and EgoZero learn visuomotor control by imitating egocentric human behavior [324, 325]. Similarly, other studies use data collected with devices such as the Apple Vision Pro to train humanoid robot policies based on natural human motion [326].

Data collection pipeline. Data collection plays a pivotal role in training VLA models. These models require large-scale, high-quality datasets, and the data acquisition pipeline must be carefully designed to ensure both efficiency and diversity. In the case of RT-1 [16], a large-scale real-robot dataset is collected using a framework that samples instructions and randomized initial states from a curated instruction set. This approach enabled the collection of demon-

Table 1: **Recent real-world robot datasets used in VLA research.** Here, *Skill* denote atomic action primitives (e.g., pick, place, reach), whereas *Task* correspond to instruction-level goals. All statistics are reported as in the original papers; the table is adapted from prior works [11, 330, 333, 335].

Name	Episodes	Skill	Task	Modality	Embodiment	Collection
QT-Opt [336]	580K	1 (Pick)	NA	RGB	KUKA LBR iiwa	Learned
MT-Opt [337]	800K	2	12	RGB, L	7 robots	Scripted, Learned
RoboNet [338]	162K	NA	NA	RGB	7 robots	Scripted
BridgeData [339]	7.2K	4	71	RGB, L	WidowX 250	Teleop
BridgeData V2 [340]	60.1K	13	NA	RGB-D, L	WidowX 250	Teleop
BC-Z [341]	26.0K	3	100	RGB, L	Google EDR	Teleop
Language Table [329]	413K	1 (Push)	NA	RGB, L	xArm	Teleop
RH20T [335]	110K	42	147	RGB-D, L, F, A	4 robots	Teleop
RT-1 [16]	130K	12	700+	RGB, L	Google EDR	Teleop
OXE [17]	1.4M	527	160,266	RGB-D, L	22 robots	Mixed
DROID [330]	76K	86	NA	RGB-D, L	Franka	Teleop
FuSe [198]	27K	2	3	RGB, L, T, A,	WidowX 250	Teleop
RoboMIND [333]	107K	38	479	RGB-D, L	4 robots	Teleop
AgiBot World [94]	1M	87	217	RGB-D, L	AgiBot G1	Teleop

strations across a broad range of tasks and environments, with human operators executing the sampled instructions to generate diverse and balanced data. In RoboTurk [327], a 6-DOF teleoperation interface was developed using an iPhone, enabling the collection of large-scale robot manipulation demonstrations via a crowdsourcing platform [328].

Furthermore, prior work [329, 330] demonstrates the effectiveness of annotating pre-collected datasets with natural language. For example, the Language Table dataset [329] collects teleoperated trajectories and subsequently adds language annotations via crowdsourcing, resulting in a large-scale dataset with approximately 600,000 language-labeled trajectories. Similarly, DROID [330] conducts distributed data collection across 18 research institutions, gathering 76,000 trajectories and 350 hours of interaction data over 564 scenes and 86 tasks, which are later annotated with natural language through a crowdsourcing platform.

However, since human annotation is costly, recent trends increasingly leverage foundation models such as VLMs to automate the annotation process. ECoT [86] and EMMA-X [331] combine object detection and gripper localization using Grounding DINO [175] and SAM [145], and high-level plan and subtask generation using Gemini 1.0 to produce automatic annotations. NILS [332] is a framework that segments long-horizon robot videos and generates language annotations without human intervention. It integrates multiple VLMs to detect keystates based on object state changes and gripper motions, and employs LLMs to generate natural language instructions. RoboMIND [333] also employs an annotation system based on Gemini [334], and demonstrates substantial performance improvements through pre-training with a VLA model.

While such methods are more cost-effective and scalable than human post-hoc annotations, they face challenges such as fine-grained scene understanding and hallucinations. Particularly in methods like ECoT that rely solely on text, inconsistencies with actual visual context are more likely to occur. Approaches grounded in visual input, such as EMMA-X, or those integrating multiple perceptual modalities, such as NILS, have proven effective in addressing these issues.

6.2 Datasets for VLA

We outline key datasets used in the pre-training of VLA models. Since the development of VLAs builds upon advances in LLMs and VLMs, a wide range of web-based datasets are leveraged. In this section, we focus specifically on datasets used for *pre-training* of VLA models, grouped into three main categories. Datasets used for post-training are typically proprietary or integrated into evaluation benchmarks such as CALVIN [342] and LIBERO [343], and are therefore excluded from this summary.

Human datasets. Collecting human data is significantly more scalable than collecting robotic data, as it does not require access to physical robots, precise calibration, or safety-critical execution environments. While third-person visual data is still used, first-person data has become particularly important for VLA pre-training because it more closely approximates the perceptual input received by real-world robots, especially those equipped with head-mounted sensors or human-like embodiments. As a result, first-person visual data is now widely adopted as a key resource for pre-training VLA models. For example, Ego4D [130] is one of the largest and most comprehensive egocentric video datasets, comprising over 3,000 hours of head-mounted RGB footage collected from more than 800 participants across 74 cities in 9 countries. Other notable examples include EPIC-KITCHENS [155, 156], which documents everyday kitchen activities, and HOI4D [344], which captures fine-grained human-object interactions. Several datasets focus specifically on manipulation tasks. OAKINK2 [345] and H2O [346] capture bimanual object manipulation using RGB-D sensors and motion capture systems. ARCTIC [347] centers on interaction with articulated objects through dexterous bimanual manipulation, while EgoPAT3D [348] focuses on human action target prediction from egocentric views.

Moreover, the advent of smart-glass-based recording devices has enabled more naturalistic and unobtrusive egocentric data collection (see Section 6.1). Notable examples include Aria Everyday Activities [323]; Ego-Exo4D [320], which integrates egocentric and exocentric perspectives; HOT3D [321], focused on fine-grained hand-object tracking; and HD-EPIC [322], which extends egocentric cooking data. These datasets are frequently used for pre-training VLA models, often via latent action prediction approaches such as LAPA [22]. Although not egocentric, large-scale video-language datasets like HowTo100M [349], Something-Something V2 [350], and Kinetics-700 [351] are also used for model pre-training and are sometimes adapted for VLA-related tasks. As VLA research increasingly employs humanoid robots and systems with human-like sensory configurations, egocentric datasets, particularly those capturing natural, goal-directed behavior, are expected to play an increasingly vital role.

Simulation datasets. Simulation environments have long been used to generate robotic datasets in a scalable, safe, and cost-effective manner. They support controlled data collection and flexible manipulation of scene configurations, making them particularly suitable for imitation learning and large-scale model pre-training. For example, RoboTurk [327] consists of task demonstrations on Sawyer robots within the MuJoCo physics engine [352], collected via remote human teleoperation over the cloud. However, collecting large-scale demonstration data in simulation, particularly via teleoperation, can still be time-consuming. To mitigate this limitation, MimicGen [353] introduces a framework for generating large-scale datasets from a small number of expert demonstrations. It decomposes demonstrations into object-centric subtasks and synthesizes new trajectories by transforming and recomposing them into novel scenes. DexMimicGen [354] extends this approach to more complex embodiments, such as dual-arm robots and multi-fingered hands.

In parallel, large-scale video world models such as COSMOS [269] have been developed to generate diverse imagined trajectories, providing rich and scalable training data for VLA models.

Although simulation played a central role in early VLA research, its dominance has declined with the increasing availability of large-scale real-world robot datasets (see the next category, which covers real robot datasets). Nonetheless, simulation remains a powerful tool for producing diverse, controllable data—particularly when real-world collection is impractical or cost-prohibitive.

Real robot datasets. Real-world robot datasets play a crucial role in the development and evaluation of VLA models. Collected on physical robot hardware, these datasets offer diverse embodiments, realistic interactions, and rich sensory inputs that are essential for training models capable of generalizing to real-world tasks. MIME [355] is one of the first large-scale robotic datasets. It contains 8.2K trajectories across 20 tasks, consisting of paired human demonstrations and kinesthetic teaching of a Baxter robot performed by humans. Concurrently, QT-Opt [336] has been introduced, comprising 580,000 grasp attempts collected over four months using seven KUKA LBR iiwa robotic arms. MT-Opt [337], an extension of QT-Opt, expands the task scope beyond grasping to support a wider range of manipulation skills. RoboNet [338] contains $162,000$ trajectories gathered across seven robot types—Sawyer, Baxter, WidowX, Franka Emika Panda, KUKA LBR iiwa, Fetch, and Google Robot. Although the trajectories are generated using random or rule-based actions rather than expert demonstrations, the dataset supports research on generalization across diverse platforms and environments. BridgeData [339] is collected via VR teleoperation using an Oculus Quest 2 and a WidowX 250 robot. It consists of $7,200$ trajectories across 10 environments and 71 tasks. An extension of this work, BridgeData V2 [340], scales the dataset to $60,000$ trajectories across 24 diverse environments. BC-Z [341] involves 12 Google Robots operated by seven human teleoperators performing over 100 manipulation tasks. Additional data are collected through policy executions with human oversight, resulting in 25,900 trajectories. Language Table [329] contains $600,000$ block manipulation trajectories (413K for real-world and 181K for simulated trajectories) paired with natural language instructions. The data are collected through long, goal-free demonstrations and annotated via crowdsourcing to support instruction-conditioned training. RH20T [335] provides multimodal data collected from four robots (Franka Emika Panda, UR5, KUKA LBR iiwa, and Flexiv Rizon) across 147 tasks and seven configurations. Unlike earlier datasets, it includes synchronized RGB-D, 6-axis force-torque, joint torque, and audio signals—supporting multimodal perception and control. RT-1 [16] comprises 130,000 real-world robotic demonstration trajectories collected over 17 months using 13 Google Robots. It serves as the foundation for the RT-series of transformer-based VLA models for real-time, instruction-conditioned behavior. Finally, Open-X Embodiment (OXE) dataset [17] unifies many of these datasets, including RT-1, BC-Z, BridgeData, and Language Table—into a standardized format using the RLDS schema [356]. Developed through a large-scale collaboration involving 21 institutions and 173 authors, OXE dataset represents one of the most comprehensive and widely adopted real-robot VLA datasets to date.

Several additional real-world robot datasets have been released to further advance VLA research. DROID [330] is a large-scale dataset comprising $76,000$ trajectories collected across 13 institutions using a standardized hardware setup.

Each participating lab used a Franka Emika Panda arm equipped with a Robotiq 2F-85 gripper, two external stereo cameras, and a wrist-mounted camera. Unlike Open X-Embodiment dataset, which aggregates data from heterogeneous robot platforms, DROID ensures consistency across environments and embodiments, making it well-suited for benchmarking. FuSe [198] provides $27,000$ multimodal trajectories collected using a WidowX 250 platform. The robot is outfitted with external cameras, a wrist-mounted camera, DIGIT tactile sensors, microphones, and an IMU, enabling rich cross-modal learning for VLA tasks. RoboMIND [333] offers $107,000$ trajectories collected from a diverse set of robot embodiments, including single-arm, dual-arm, humanoid, and dexterous-hand configurations. The dataset emphasizes diversity in morphology and manipulation strategies, supporting research in generalization and transfer. AgiBot World Dataset [94] is a massive-scale dataset comprising 1 million trajectories collected using over 100 AgiBot G1 robots. Its unprecedented scale enables training of large VLA models under highly diverse conditions. In addition to these major releases, several task-specific or platform-specific datasets have been introduced, including Task-Agnostic Robot Play [357, 358], Jaco Play [359], Cable Routing [360], Berkeley Autolab UR5 [361], TOTO [362], and RoboSet [363]. Navigation-focused VLA datasets have also emerged, such as SACSoN [364], SCAND [365], RECON [366], and BDD100K [367], which support instruction-following and goal-directed behaviors in mobile platforms. Finally, specialized datasets such as RoboVQA [368] target robot-specific question answering, further broadening the scope of VLA applications beyond manipulation and navigation.

6.3 Data Augmentation for VLA

Given the high cost of collecting datasets, various data augmentation methods have been developed to expand existing datasets. These approaches span multiple modalities, including vision, language, and action.

Vision augmentation. In most computer vision tasks, augmentation techniques such as rotation, cropping, and scaling are commonly used to improve generalization. However, in robotics, where the robot's embodiment and its spatial relationship to the camera are critical, such transformations can distort these relationships and negatively affect performance. To address this, recent methods have proposed using image generation models, such as Stable Diffusion [137], to perform embodiment-aware augmentations. CACTI [369] leverages Stable Diffusion to modify a specific region of images to augment a small, yet high-quality dataset. GenAug [370] introduces more sophisticated visual augmentation by leveraging Stable Diffusion to apply three types of transformations: altering object textures, inserting task-irrelevant distractors, and modifying backgrounds. These augmentations aim to improve policy robustness by increasing visual diversity while preserving task-relevant semantics. ROSIE [371] builds on CACTI and GenAug by using an LLM, OWL-ViT [144], and Imagen Editor [372] to automatically identify and modify masked regions based on text prompts, enabling controlled edits to target objects, backgrounds, or the insertion of new objects. The augmented data is used to train RT-1 [16]. DreamGen [110] utilizes a video world model to generate diverse visual variations, paired with an inverse dynamics model (IDM) to infer the corresponding actions. This combination enables the synthesis of training data, facilitating policy learning in novel environments and enhancing generalization. In contrast, MOO [57] forgoes explicit visual augmentation and instead disentangles object and skill representations using a vision-language model (VLM), allowing policies to generalize to unseen object-skill combinations from limited data. It addresses visual variability implicitly by leveraging the broad generalization capabilities of pre-trained VLMs. Moreover, BYOVLA [373] extracts and inpaints task-irrelevant regions in image observations during runtime, aiming to enhance robustness against visual distractions.

Language augmentation. DIAL [374] starts with a small, manually labeled seed set of trajectory-instruction pairs. A VLM is trained on this seed set to compute similarity between trajectories and instructions. Simultaneously, an LLM generates diverse paraphrases of the seed instructions, forming a large pool of candidates. These are then matched to the remaining unlabeled trajectories using the trained VLM, and the top-k most similar instructions are assigned. The resulting dataset is used to train RT-1 [16].

Action augmentation. Since actions are directly tied to the robot's physical behavior and embodiment, augmenting action data is generally challenging. A common approach to address this challenge is dataset expansion through interactive methods such as DAgger [375], which iteratively collects expert actions in states visited by the learned policy. Similarly, CCIL [376] generates corrective data when a policy encounters out-of-distribution states by learning a locally smooth dynamics model. It synthesizes actions that guide the robot from novel states back to expert-visited ones, and the resulting corrective data is combined with the original demonstrations to refine the policy.

7 Review of Real-World Robot Applications

In this section, we summarize key practical aspects of VLA research, including the types of robots used, data collection methodologies, publicly available datasets and augmentation techniques, and the evaluation protocols applied to assess model performance.

7.1 Robot for VLA

In this section, we present an overview of the types of robots commonly employed in VLA research.

Manipulator. Robotic manipulators are the most commonly used robots in VLA research, encompassing both single-arm and dual-arm configurations. Single-arm robots used in the prior works reviewed in this survey include: Franka Emika Panda, Franka Research 3, UR5, UR5e, UR3, UR3e, UR10, Kinova Gen3, Kinova Jaco 2, Sawyer, KUKA LBR iiwa 14, UFactory xArm, DENSO Cobotta, FANUC LR Mate 200iD, Realman RM65-B, Realman RM75-6F, AgileX PiPER, Unitree Z1 Pro, Dobot, Flexiv Rizon, AIRBOT Play, ARX, DLR SARA [377], WidowX 250 6DoF, ViperX 300 6DoF, SO-100/101, and PAMY2 [378]. These manipulators typically feature 5, 6, or 7 degrees of freedom (DoFs). The joint configurations and link lengths vary across these manipulators. PAMY2 uses pneumatic actuation, reflecting the diversity of robotic embodiments. In addition, several systems (e.g., AgileX PiPER, ARX, Franka Emika Panda, UFactory xArm, UR5e, AIRBOT Play, ALOHA [301], and ALOHA2 [303] adopt a bimanual configuration by placing two arms side by side. WidowX, ViperX, ALOHA, SO-100/101, and PAMY2 are fully open-source in hardware, allowing researchers to flexibly modify or extend their physical embodiment. These manipulators are used to perform a wide range of tasks, including object grasping and relocation, assembly, manipulation of deformable objects, and peg-in-hole insertion.

Hand / Gripper. This category refers to the hands and grippers that serve as end-effectors mounted on the manipulators described above. Hands used in prior works in VLA include the ROBOTERA Xhand, PSYONIC Ability Hand, Inspire Robots RH56, Shadow Hand, PsiBot G0-R, Robotiq 2F-85/140, LEAP Hand, and UMI. These vary in design: the LEAP Hand [379] has four fingers; the Robotiq Gripper and UMI [313] are two-fingered; the others are five-fingered. Some systems also use suction cups or task-specific grippers, as in Shake-VLA [380]. Platforms such as ALOHA, ARX, and PiPER typically include two-fingered grippers by default. The LEAP Hand and UMI are open-source, allowing easy hardware modification. While two-fingered grippers are suited for grasping, four- and five-fingered hands enable tool use and in-hand manipulation.

Mobile robot. Mobile robots in VLA research include both wheeled platforms and mobile manipulators that combine robotic arms with mobile bases. Jackal and TurtleBot 2 are examples of systems that rely exclusively on wheeled locomotion and do not incorporate manipulation capabilities. In contrast, mobile manipulators exhibit diverse configurations, including single-arm platforms such as Hello Stretch, Google Robot, and LoCoBot, as well as dual-arm systems like Mobile ALOHA, PR2, Fibocom, and AgiBot G1. LoCoBot and TurtleBot 2 are also notable for their fully open-source hardware, which facilitates embodiment customization and experimentation. Mobile platforms enable locomotion and environmental interaction capabilities beyond those afforded by stationary arms or grippers, supporting tasks that involve navigation and dynamic scene engagement. Some models, such as RT-1, are capable of performing navigation and manipulation concurrently.

Quadruped robot. Quadruped robots, characterized by their animal-like locomotion, have been increasingly considered in VLA research due to their ability to navigate unstructured and uneven environments. Unitree A1, Go1, Go2, B1, Boston Dynamics Spot, and ANYmal are frequently used. These are all commercially available systems capable of traversing complex terrain using RL-based control policies. These platforms not only provide locomotion but can also be equipped with manipulators to support a wide range of manipulation tasks.

Humanoid robot. Humanoid robots, characterized by body structures resembling those of humans, represent another category of platforms explored in VLA research. In prior works, Fourier GR-1, Unitree G1, Unitree H1, and Booster T1 are often used. These systems typically possess two legs, two arms, and five-fingered hands attached to their end effectors. Their human-like morphology makes them well-suited for operating in spaces designed for humans and facilitates compatibility with VLAs trained on human motion datasets.

7.2 Evaluation for VLA

Evaluation metrics for VLA models remain poorly defined, particularly in real-world settings. Assessing generalization on physical robots is challenging due to differences in embodiment, safety concerns, and limited reproducibility. Consequently, most evaluations are conducted in simulation, where standardized environments and benchmarks facilitate fair comparisons across methods. Below, we introduce representative simulation environments and their variants commonly used for evaluating and comparing VLA models.

MuJoCo. Several simulation environments have been developed on top of MuJoCo [352] to support research in robotic manipulation. For example, robosuite [381], a modular simulation framework in which robots, arenas, and task objects are composed using MJCF files, provides 11 manipulation tasks.

Table 2: **Benchmarks for VLA evaluation.** This table shows various simulation environments used for evaluating VLA models with their key characteristics. Task types include Navigation (Nav), Manipulation (Manip), and Whole-Body Control (WBC). Observation modalities include RGB-D (RGB + Depth), S (Semantic segmentation), and PC (Point Cloud). The Scenes/Objects column indicates the number of available scenes and objects respectively.

Name	Task	Scenes/Objects	Observation	Physics	Built Upon	Description
robosuite [381]	Manip	NA / 10	RGB-D, S	MuJoCo	NA	Modular framework, 11 tasks
robomimic [382]	Manip	NA / NA	RGB	MuJoCo	robosuite	Offline learning, 8 tasks
RoboCasa [383]	Manip	120 / 2.5K	RGB	MuJoCo	robosuite	100 kitchen tasks, photorealistic
LIBERO [343]	Manip	NA / NA	RGB	MuJoCo	robosuite	130 tasks in 4 task suites
Meta-World [384]	Manip	1 / 80	Pose	MuJoCo	NA	50 Manip tasks for Meta-RL
LeVERB-Bench [385]	Nav, WBC	4 / NA	RGB	PhysX	Isaac Sim	Humanoid control
ManiSkill [386]	Manip	NA / 162	RGB-D, PC, S	PhysX	SAPIEN	4 tasks, 36K demos
ManiSkill 2 [387]	Manip	NA / 2.1K	RGB-D, PC	PhysX	ManiSkill	Extended task diversity
ManiSkill 3 [388]	Nav, Manip, WBC	NA / NA	RGB-D, PC, S	PhysX	ManiSkill 2	GPU-parallelized simulation
ManiSkill-HAB [389]	Manip	105 / 92	RGB-D	PhysX	ManiSkill 3, Habitat 2.0	HAB tasks from Habitat 2.0
RoboTwin [390, 391]	Manip	NA / 731	RGB-D	PhysX	SAPIEN	Dual-arm tasks
Ravens [26]	Manip	NA / NA	RGB-D	PyBullet	NA	10 tabletop tasks
VIMA-BENCH [31]	Manip	NA / 29	RGB, S	PyBullet	Ravens	17 multimodal prompt tasks
LoHoRavens [392]	Manip	1 / 3	RGB-D	PyBullet	Ravens	Long-horizon planning
CALVIN [342]	Manip	4 / 7	RGB-D	PyBullet	NA	Long-horizon lang-cond tasks
Habitat [393]	Nav	185 / NA	RGB-D, S	Bullet	NA	Fast, Nav only
Habitat 2.0 [394]	Nav, Manip	105 / 92	RGB-D	Bullet	Habitat	Mobile manipulation (HAB)
Habitat 3.0 [395]	Nav, Manip	211 / 18K	RGB-D	Bullet	Habitat 2.0	Human avatars support
RLBench [396]	Manip	1 / 28	RGB-D, S	PyBullet	V-REP	Tiered task difficulty
THE COLOSSEUM [397]	Manip	1 / 107	RGB-D	PyBullet	RLBench	20 tasks, 14 env variations
AI2-THOR [398]	Nav, Manip	NA / 118	RGB-D, S	Unity	NA	Object states, task planning
CHORES [399]	Nav	191K / 40K	RGB	Unity	AI2-THOR	Shortest-path planning
SIMPLER [400]	Manip	4 / 17	RGB	PhysX	SAPIEN, Isaac Sim	Real-to-sim evaluation
RoboArena [401]	Manip	NA / NA	RGB	Real	NA	Distributed real-world evaluation

Building on robosuite, robomimic [382] introduces a systematic benchmark for evaluating learning from demonstrations in robotic manipulation. The robomimic benchmark includes 8 tasks performed using a Franka Emika Panda robot.

RoboCasa [383] further extends robosuite by incorporating large-scale, photorealistic scenes that span 100 tasks across a variety of robot platforms, enabling broader generalization and transfer learning studies. Currently, the most widely used benchmark for evaluating VLA models is LIBERO [402], which is designed for language-conditioned manipulation tasks. It provides 4 task suites comprising a total of 130 tasks, all executed by a Franka Emika Panda robot: LIBERO-SPATIAL focuses on spatial reasoning between objects, LIBERO-OBJECT targets object category recognition, LIBERO-GOAL evaluates understanding of object manipulation goals, and LIBERO-100 integrates the three previous suites to assess compositional generalization. Furthermore, Meta-World [384] is another simulation environment built on MuJoCo, designed to evaluate multi-task and meta-reinforcement learning. It includes 50 distinct tasks performed using a Sawyer robotic arm, enabling evaluation of generalization across diverse manipulation skills.

PhysX. IsaacLab [403] is a GPU-accelerated framework built on IsaacSim, which employs PhysX as its underlying physics engine. It provides a comprehensive suite of tools for robot learning, including a diverse set of robots, environments, and sensors, along with photorealistic rendering capabilities. LeVERB-Bench [385], also built on IsaacSim, focuses on full-body humanoid control and includes 154 vision-language tasks and 460 language-only tasks.

Moreover, ManiSkill [386–388], built on the SAPIEN simulation platform [404], whose underlying physics engine is also based on PhysX, serves as a comprehensive benchmark for learning object manipulation skills from 3D visual input. It includes a wide range of tasks involving articulated and deformable objects, mobility, and diverse robot embodiments, and provides large-scale demonstration data with support for efficient, high-quality simulation. ManiSkill-HAB [389] is a benchmark focused on object rearrangement tasks that follow the Home Assistant Benchmark (HAB) introduced in Habitat 2.0 [394]. In addition, several other benchmarks have been developed on SAPIEN, such as RoboCAS [405], which evaluates robotic manipulation in complex object arrangement environments, and DexArt [406], which focuses on manipulation of articulated objects using multi-fingered hands. More recently, RoboTwin [390, 391] has been proposed as a benchmark for dual-arm manipulation, offering 50 tasks, 731 objects, and 5 distinct embodiments.

Bullet. Ravens [26] is a benchmark of 10 tabletop manipulation tasks implemented using PyBullet [407]. VIMA-BENCH [31] extends this benchmark with 17 tasks that allow multi-modal prompt-based task specification. LoHoRavens [392] is another extension that evaluates long-horizon planning capabilities in tabletop manipulation scenarios. Moreover, CALVIN [342] provides a simulation and benchmark for long-horizon manipulation based on natural language instructions, which includes 34 manipulation tasks performed by a Franka Emika Panda robot. In addition, Habitat [393–395] is a simulation framework primarily developed by Meta. Habitat 1.0 [393] provides a simulation platform specialized for visual navigation tasks. Habitat 2.0 [394] extends this to mobile manipulation tasks and introduces the Home Assistant Benchmark (HAB). Further, Habitat 3.0 [395] expands the framework to support not only robots but also human avatars.

V-REP. RLBench [396] is the first large-scale benchmark for imitation and reinforcement learning, built using V-REP [408] and PyRep [409]. It contains 100 manipulation tasks using the Panda robot. THE COLOSSEUM [397], built on top of RLBench, is a benchmark designed to systematically evaluate the generalization capabilities of robotic manipulation policies under environment variations. THE COLOSSEUM includes 20 manipulation tasks with 14 types of environment perturbations.

Unity. AI2-THOR is a photorealistic, interactive 3D simulation environment built on the Unity engine, offering four task suites, such as iTHOR, RoboTHOR, ProcTHOR-10K, and ArchitecTHOR [398, 410, 411], that collectively encompass a diverse range of indoor environments. Moreover, SPOC [399] introduces CHORES, an extension of AI2-THOR designed as a benchmark for shortest-path planning in navigation tasks.

Miscellaneous. While not strictly simulation-based benchmarks, several studies have proposed evaluation protocols to assess the capabilities of VLA models. VLATest [412] systematically evaluates the impact of various factors on VLA model performance, including the number of confounding objects, lighting conditions, camera poses, unseen objects, and mutations in task instructions. Moreover, several works aim to improve robustness against adversarial attacks [413, 414] and enhance interpretability by probing the latent representations of VLA models to uncover symbolic structures corresponding to object properties, spatial relations, and action states [415].

Toward realistic and scalable evaluation for VLA. There is increasing emphasis on evaluation under conditions that closely resemble the real world, leading to the development of both realistic simulation benchmarks and scalable systems for distributed real-world evaluation of VLA models. SIMPLER [400] enables the evaluation of policies trained on real-world data within simulation by minimizing visual and control domain gaps, achieving high correlation between simulation and real-world performance. RoboArena [401] is a distributed framework for large-scale, fair, and reliable evaluation of VLA models in the real world. It conducts pairwise comparisons across a network of robots deployed at seven universities, with results aggregated by a central server to produce global rankings. This system is built on the DROID platform.

7.3 Real-world Applications

This section provides concrete examples of how the previously introduced robotic platforms, including manipulators, hands, mobile robots, quadrupeds, and humanoids, are employed in the development and evaluation of VLA models.

Manipulator. Manipulators represent the most widely used robotic platforms in VLA research. They are employed across a diverse set of tasks, including object grasping and relocation, assembly, deformable object manipulation, and peg-in-hole insertion. Both single-arm and more complex dual-arm robots are commonly utilized, enabling a broader range of dexterous manipulation tasks. Notable demonstrations in this domain include Shake-VLA [380], which performs cocktail mixing using dual-arm coordination, and RoboNurse-VLA [205], which automates surgical instrument handovers in clinical environments.

Hand / Gripper. Hands and grippers, commonly used as end-effectors on manipulators, enable a wide range of manipulation tasks. Two-fingered grippers are particularly well suited for object grasping, while more dexterous four- and five-fingered robotic hands facilitate tool use and in-hand manipulation. For instance, GraspVLA [100] develops a VLA model for object grasping using a two-fingered gripper. In contrast, DexGraspVLA [75] leverages a multi-fingered robotic hand to construct a VLA model capable of performing more delicate and precise grasping tasks.

Mobile robot. Mobile robots are primarily utilized in VLA models for navigation-related tasks [416]. Beyond navigation, models such as RT-1 [16] are capable of generating both arm and base motions for mobile manipulators, robots that integrate a mobile base with a robotic arm. The VLA framework has also been extended to other mobile domains. For instance, aerial robots such as the DJI Tello are used in UAV-based VLA research, with works including UAV-VLA [288], RaceVLA [417], and CognitiveDrone [418] focusing on autonomous flight. Similarly, VLA applications in autonomous driving have been explored in OpenDriveVLA [226], ORION [174], CoVLA [89], and OccLLaMA [225]. These developments demonstrate the adaptability of VLA systems across a diverse range of mobile robotic platforms.

Quadruped robot. Quadruped robots enable more diverse and versatile navigation compared to wheeled mobile robots due to their ability to traverse uneven, unstructured, and dynamic terrains. Several prior works, including TrackVLA [105,243], NaVILA [243], and CrossFormer [419], successfully demonstrate robust navigation capabilities, including deployment in the wild. Furthermore, Track2Act [420] and VidBot [159] utilize Boston Dynamics Spot equipped with a manipulator for integrated navigation and manipulation in home environments. SLIM [265] similarly employs a Unitree Go1 equipped with a mounted WidowX 250 arm to perform multimodal tasks, such as grasping objects from the ground while navigating uneven terrain.

Humanoid robot. Humanoids have gained significant attention in VLA research, because their human-like morphology offers practical advantages for real-world deployment, as most environments, tools, and interfaces are designed for human use, making task transfer and embodiment alignment more straightforward. NaVILA [243] demonstrates robust locomotion capabilities in tightly controlled laboratory settings. In contrast, EgoVLA [421] and GO-1 [94] focus on manipulation tasks commonly encountered in household environments, including picking, placing, pouring, and folding.

8 Recommendations for Practitioners

Drawing on insights from recent VLA research, this section provides actionable recommendations for practitioners seeking to design, train, and deploy VLA models in real-world robotic systems. We highlight practical strategies across data collection, architecture selection, and model adaptation.

Prioritize diverse and high-quality datasets. Robust generalization across tasks, objects, and embodiments relies on training with large-scale, high-quality datasets that encompass vision, language, and action modalities. Practitioners should aim to collect or utilize datasets that offer broad task coverage, environmental variability, and embodiment diversity. Such diversity is essential for improving the robustness and transferability of VLA policies.

Prefer continuous action generation via generative methods. While it is increasingly well established in recent literature, generating continuous actions, rather than relying on discretized tokens, remains critical for achieving smooth and precise robot behavior. Practitioners are encouraged to adopt generative approaches such as diffusion or flow matching to enable high-fidelity control in real-world settings.

Try gradient insulation during pre-training. Allowing gradients from randomly initialized action heads to propagate into pre-trained VLM backbones can degrade the quality of learned representations that already capture common-sense knowledge. To stabilize training and preserve the semantic knowledge in the backbone, practitioners are encouraged to freeze the backbone or apply gradient insulation mechanisms. This approach has been shown to improve both training efficiency and final performance.

Begin with lightweight adaptation methods. Full fine-tuning of large VLA models is often computationally prohibitive. As a first step, practitioners, who do not have access to a GPU cluster, can fine-tune only the action head while keeping the backbone frozen. Alternatively, methods such as LoRA enable parameter-efficient fine-tuning, offering a favorable trade-off between performance and resource consumption.

Incorporate world models or latent action learning for scalability. In scenarios involving humanoid robots, incorporating human video data during pre-training can be particularly advantageous due to the similarity in embodiment. However, as such datasets typically lack explicit action annotations, it is beneficial to learn latent action representations that can be used as surrogate action targets during pre-training. In addition, the predictive capabilities of world models can support more effective planning and reasoning, especially in manipulation tasks. By anticipating future observations, world models facilitate better long-horizon control and multimodal grounding, as demonstrated in prior work such as FLARE [139].

Embrace multi-task learning to enhance representations for action generation. While VLMs pre-trained on web-scale data offer strong semantic grounding, their representations are not always directly suited for downstream control. Incorporating auxiliary tasks such as affordance estimation, keypoint detection, future state prediction, and segmentation for a target object encourages the model to learn representations that are better aligned with the requirements of action generation. These tasks support spatial reasoning, temporal prediction, and physical interaction modeling, ultimately improving the model's ability to translate perception into effective control.

9 Future Research Direction

9.1 Data Modality

While several prior works have attempted to integrate additional modalities such as audio, tactile sensing, and 3D point clouds into VLA models, collecting large-scale datasets with such modalities remains a significant challenge. In particular, tactile sensing poses serious difficulties due to the diversity of sensor types, data formats, and hardware configurations. The lack of standardization across robotic platforms further complicates multimodal data collection and integration. Although tactile feedback is likely essential for achieving human-level dexterous manipulation, current tactile sensors vary widely in design and are not yet widely adopted. Therefore, unifying sensor configurations is critical to enabling scalable, multimodal VLA systems.

9.2 Reasoning

Reasoning is a particularly important capability for solving long-horizon tasks in VLA systems. Beyond anticipating future events based on current observations, effective reasoning requires the ability to retain relevant information over time and retrieve it when needed. This involves maintaining a form of memory and selectively attending to key information that supports decision-making across temporally extended tasks. For example, in mobile robot manipulation, a typical task may involve first locating a shelf, then navigating to a different location to pick up a cup, and finally returning to place the cup on the shelf. In such cases, the robot must remember the location of the shelf encountered earlier and retrieve that information at the appropriate time. This type of temporal abstraction and memory-based retrieval is essential for robust reasoning and planning in real-world scenarios. Enhancing these capabilities is likely to be a key direction for future research in VLA systems, particularly as tasks grow in complexity and duration.

9.3 Continual Learning

A fundamental limitation of current VLA systems is their inability to learn beyond their initial training phase. Once trained offline, these models are typically frozen and do not adapt to new situations. Unlike humans, who continuously learn from ongoing experience, VLA systems remain fixed, making them vulnerable when faced with novel or out-of-distribution scenarios. In such cases, the robot may fail to act appropriately. To overcome this limitation, enabling online or continual learning will be essential. By incrementally updating their internal representations and policies based on new data, VLA systems could better adapt to diverse environments. However, this capability introduces several challenges, including catastrophic forgetting and safety concerns related to deploying untested updates in real-world settings. Despite these difficulties, continual learning remains a promising direction for future VLA research. Approaches such as reinforcement learning from human feedback (RLHF) and active learning inspired by cognitive development may offer viable pathways toward building adaptive, lifelong-learning VLA systems capable of operating safely and effectively in the real world.

9.4 Reinforcement Learning

While several prior studies [254, 255, 261] have explored the use of RL to fine-tune vision-language-action (VLA) models, these efforts have predominantly focused on evaluation in simulated environments. This is largely due to the substantial number of samples required for RL and the risk of unsafe behavior during real-world exploration. As a result, fine-tuning VLA models within a learned world model presents a promising research direction, offering a safer and more sample-efficient alternative. In addition, real-to-sim techniques allow the construction of digital twin environments in which VLA models can be fine-tuned using RL. However, challenges remain in accurately identifying physical parameters and reconstructing scenes, the latter of which often requires multi-view observations [422]. Overall, we posit that advances in world modeling and real-to-sim transfer may enable scalable and safe fine-tuning of VLA models through RL.

9.5 Safety

While VLA models perform well on manipulation tasks in controlled settings, their deployment in unstructured environments poses significant safety challenges. Current systems often lack mechanisms to detect and avoid unexpected human presence in the workspace, increasing the risk of collisions. Although collecting demonstrations of such edge cases is possible, doing so via teleoperation remains risky, as the robot may not respond safely in real time. This underscores the need to integrate VLA with model-based control approaches, which offer predictive reasoning in safety-critical situations. We argue that improving the safety of VLA systems requires hybrid architectures that combine the generalization capabilities of learned policies with the reliability of model-based controllers.

9.6 Failure Detection and Recovery

In real-world environments, unexpected failures are often unavoidable. However, most current VLA systems lack mechanisms for detecting such failures or responding appropriately. Failures are typically treated as terminal events, with no recovery or re-planning strategies in place. To enable reliable deployment in practical applications, it is essential for VLA systems to detect failures and adapt their behavior accordingly. Several recent works have begun to address this gap. SAFE [423] leverages intermediate representations within VLA models to identify failure events during execution. Agentic Robot [424] uses a vision-language model (VLM) to detect failures, execute predefined recovery behaviors, and then re-plan the task. A more robust solution is proposed in LoHoVLA [245], which employs a hierarchical architecture. Upon detecting a failure, the system regenerates the current action; if the same failure is detected multiple times, it escalates the response by re-generating the higher-level subtask, thus enhancing overall robustness. FOREWARN [283] introduces a predictive planning mechanism by sampling a large number of action sequences from the policy, clustering them into six behavioral modes, and using the DreamerV3 world model [425] to simulate future states. The most promising behavioral mode is then selected based on these predictions. As VLA systems are increasingly applied to long-horizon and open-ended tasks, the ability to detect failures and recover through adaptive re-planning will be critical for achieving robustness and reliability in real-world deployment.

9.7 Evaluation

While various VLAs with different architectures, modalities, and training methods have been proposed, it remains unclear which approaches yield the most effective performance. This ambiguity largely stems from the lack of a statistically rigorous evaluation. As demonstrated in LBM [314], it is crucial to conduct evaluations under controlled and comparable conditions, with a sufficient number of evaluation trials and appropriate statistical analysis (e.g. confidence intervals) to ensure whether observed performance differences are statistically significant.

9.8 Applications

VLA systems have potential applications across a wide range of domains, including healthcare, assistive technologies, industrial automation, and autonomous driving. However, despite this breadth of applicability, VLA models have not yet reached the level of performance or reliability required for practical deployment. Most existing systems operate only within constrained, predefined environments and still fall short of human-level capabilities in terms of robustness and adaptability.

As the field increasingly prioritizes real-world use cases, there will likely be growing attention to issues such as safety, reliability, and operational efficiency, key factors that must be addressed to enable the successful deployment of VLA systems in practical applications.

10 Conclusion

This survey provides a comprehensive review of Vision-Language-Action (VLA) models for robotics, tracing their evolution from early CNN-based approaches to sophisticated multimodal architectures integrating diffusion models and latent action representations. We have examined the fundamental challenges, architectural innovations, training methodologies, and real-world applications that define the current landscape of VLA research.

Our analysis reveals several key insights: (1) the critical role of large-scale datasets and pre-trained foundation models in enabling generalization, (2) the emergence of hierarchical architectures that separate high-level reasoning from low-level control, (3) the growing importance of multimodal inputs beyond vision and language, and (4) the persistent challenges in sim-to-real transfer and embodiment generalization. The field has reached a critical inflection point at which recent advances in foundation models, in conjunction with improved data collection protocols and refined training methodologies, are anticipated to facilitate the development of robotic systems with improved generalization and capability. The incorporation of world models, affordance-based reasoning, and RL is expected to underpin the next generation of VLA models, enabling continuous learning, sophisticated task reasoning, and robust adaptation across diverse and unstructured real-world environments.

References

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela

Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[2] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023.

[4] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian

Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[5] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Hao-Shu Fang, Shibo Zhao, Shayegan Omidshafiei, Dong-Ki Kim, Ali akbar Agha-mohammadi, Katia Sycara, Matthew Johnson-Roberson, Dhruv Batra, Xiaolong Wang, Sebastian Scherer, Chen Wang, Zsolt Kira, Fei Xia, and Yonatan Bisk. Toward general-purpose robots via foundation models: A survey and meta-analysis. 2023.

[6] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo Guo, Chris Paxton, and Andy Zeng. Real-world robot applications of foundation models: A review. *Advanced Robotics*, Vol. 38, No. 18, pp. 1232–1254, 2024.

[7] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran

Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, Vol. 44, No. 5, pp. 701–739, 2025.

[8] brian ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning (CoRL)*, Vol. 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 14–18 Dec 2023.

[9] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, 2023.

[10] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Vol. 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023.

[11] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

[12] Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

[13] Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, Yuanpei Chen, and Yaodong Yang. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025.

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[15] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning (CoRL)*, Vol. 164 of *Proceedings of Machine Learning Research*, pp. 894–906. PMLR, 08–11 Nov 2022.

[16] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[17] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang

Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi Jim Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick Tree Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024.

[18] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[19] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

[20] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

[21] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[22] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations (ICLR)*, 2025.

[23] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[24] Johan Bjorck, Fernando Casta neda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[26] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In Jens Kober, Fabio Ramos, and Claire Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning (CoRL)*, Vol. 155 of *Proceedings of Machine Learning Research*, pp. 726–747. PMLR, 16–18 Nov 2021.

[27] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. Curran Associates, Inc., 2017.

[29] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[31] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022. PMLR, 23–29 Jul 2023.

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67, 2020.

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019.

[35] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[36] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[37] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34, pp. 12786–12797. Curran Associates, Inc., 2021.

[38] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 23–29 Jul 2023.

[39] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14432–14444, June 2024.

[40] Priya Sundaresan, Quan Vuong, Jiayuan Gu, Peng Xu, Ted Xiao, Sean Kirmani, Tianhe Yu, Michael Stark, Ajinkya Jain, Karol Hausman, Dorsa Sadigh, Jeannette Bohg, and Stefan Schaal. Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 70–96. PMLR, 06–09 Nov 2025.

[41] Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, Priya Sundaresan, Peng Xu, Hao Su, Karol Hausman, Chelsea Finn, Quan Vuong, and Ted Xiao. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. In *International Conference on Learning Representations (ICLR)*, 2024.

[42] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quan Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[43] Isabel Leal, Krzysztof Choromanski, Deepali Jain, Avinava Dubey, Jake Varley, Michael Ryoo, Yao Lu, Frederick Liu, Vikas Sindhwani, Quan Vuong, Tamas Sarlos, Ken Oslund, Karol Hausman, and Kanishka Rao. Sara-rt: Scaling up robotics transformers with self-adaptive robust attention. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6920–6927, 2024.

[44] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Sean Kirmani, Isabel Leal, Edward Lee, Sergey Levine, Yao Lu, Isabel Leal, Sharath Maddineni, Kanishka Rao, Dorsa Sadigh, Pannag Sanketi, Pierre Sermanet, Quan Vuong, Stefan Welker, Fei Xia, Ted Xiao, Peng Xu, Steve Xu, and Zhuo Xu. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024.

[45] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vol. 235 of *Proceedings of Machine Learning Research*, pp. 23123–23144. PMLR, 21–27 Jul 2024.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2024.

[47] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

[48] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

[50] Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025.

[51] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[52] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[53] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. Curran Associates, Inc., 2017.

[54] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024.

[55] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[56] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

[57] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, Chelsea Finn, and Karol Hausman. Open-world object manipulation using pre-trained vision-language models. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Vol. 229 of *Proceedings of Machine Learning Research*, pp. 3397–3417. PMLR, 06–09 Nov 2023.

[58] Priya Sundaresan, Quan Vuong, Jiayuan Gu, Peng Xu, Ted Xiao, Sean Kirmani, Tianhe Yu, Michael Stark, Ajinkya Jain, Karol Hausman, Dorsa Sadigh, Jeannette Bohg, and Stefan Schaal. Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches. *arXiv preprint arXiv:2403.02709*, 2024.

[59] Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, Priya Sundaresan, Peng Xu, Hao Su, Karol Hausman, Chelsea Finn, Quan Vuong, and Ted Xiao. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977*, 2023.

[60] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X. Lee, Maria Bauza Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Fernandes Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott Reed, Sergio Gómez Colmenarejo, Jonathan Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Thomas Rothörl, Jose Enrique Chen, Yusuf Aytar, David Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2024.

[61] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.

[62] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26439–26455, June 2024.

[63] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z. Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[64] Siddhant Haldar, Zhuoran Peng, and Lerrel Pinto. Baku: An efficient transformer for multi-task policy learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37, pp. 141208–141239. Curran Associates, Inc., 2024.

[65] Yueen Ma, Dafeng Chi, Shiguang Wu, Yuecheng Liu, Yuzheng Zhuang, Jianye Hao, and Irwin King. Actra: Optimized transformer architecture for vision-language-action models in robot learning. *arXiv preprint arXiv:2408.01147*, 2024.

[66] Xincheng Pang, Wenke Xia, Zhigang Wang, Bin Zhao, Di Hu, Dong Wang, and Xuelong Li. Depth helps: Improving pre-trained rgb-based policy with depth information injection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7251–7256, 2024.

[67] Ria Doshi, Homer Rich Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 496–512. PMLR, 06–09 Nov 2025.

[68] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[69] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 63–70, 2024.

[70] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Tinyvla: Toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 10, No. 4, pp. 3988–3995, 2025.

[71] Sicheng Wang, Sheng Liu, Weiheng Wang, Jianhua Shan, and Bin Fang. Robobert: An end-to-end multimodal robotic manipulation model. *arXiv preprint arXiv:2502.07837*, 2025.

[72] Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27661–27672, June 2025.

[73] Weiyu Liu, Yilun Du, Tucker Hermans, Sonia Chernova, and Chris Paxton. Structdiffusion: Language-guided creation of physically-valid structures using unseen objects. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[74] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal Diffusion Transformer: Learning Versatile Behavior from Multimodal Goals. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[75] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Yitao Liang, Yaodong Yang, and Yuanpei Chen. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*, 2025.

[76] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[77] Rujia Yang, Geng Chen, Chuan Wen, and Yang Gao. Fp3: A 3d foundation policy for robotic manipulation. *arXiv preprint arXiv:2503.08950*, 2025.

[78] Yuanqi Yao, Siao Liu, Haoming Song, Delin Qu, Qizhi Chen, Yan Ding, Bin Zhao, Zhigang Wang, Xuelong Li, and Dong Wang. Think small, act big: Primitive prompt learning for lifelong robot manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22573–22583, June 2025.

[79] Yuyin Yang, Zetao Cai, Yang Tian, Jia Zeng, and Jiangmiao Pang. Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation. *arXiv preprint arXiv:2504.17784*, 2025.

[80] Zhi Hou, Tianyi Zhang, Yuwen Xiong, Hengjun Pu, Chengyang Zhao, Ronglei Tong, Yu Qiao, Jifeng Dai, and Yuntao Chen. Diffusion transformer policy. *arXiv preprint arXiv:2410.15959*, 2024.

[81] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vol. 235 of *Proceedings of Machine Learning Research*, pp. 20413–20451. PMLR, 21–27 Jul 2024.

[82] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

[83] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37, pp. 40085–40110. Curran Associates, Inc., 2024.

[84] Pengxiang Ding, Han Zhao, Wenjie Zhang, Wenxuan Song, Min Zhang, Siteng Huang, Ningxi Yang, and Donglin Wang. Quar-vla: Vision-language-action model for quadruped robots. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pp. 352–367, Cham, 2025. Springer Nature Switzerland.

[85] Xiang Li, Cristina Mata, Jongwoo Park, Kumara Kahatapitiya, Yoo Sung Jang, Jinghuan Shang, Kanchana Ranasinghe, Ryan Burgert, Mu Cai, Yong Jae Lee, and Michael S. Ryoo. Llara: Supercharging robot learning data for vision-language policy. In *International Conference on Learning Representations (ICLR)*, 2025.

[86] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 3157–3181. PMLR, 06–09 Nov 2025.

[87] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[88] Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024.

[89] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 1933–1943, February 2025.

[90] Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, and Feifei Feng. Diffusion-vla: Generalizable and interpretable robot foundation model via self-generated reasoning. *arXiv preprint arXiv:2412.03293*, 2024.

[91] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

[92] Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Ran Cheng, Yaxin Peng, Chaomin Shen, and Feifei Feng. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. *arXiv preprint arXiv:2502.14420*, 2025.

[93] Minjie Zhu, Yichen Zhu, Jinming Li, Zhongyi Zhou, Junjie Wen, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. Objectvla: End-to-end open-world object manipulation without demonstration. *arXiv preprint arXiv:2502.19250*, 2025.

[94] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[95] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511*, 2025.

[96] Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Yuan Du, and Shanghang Zhang. Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. *arXiv preprint arXiv:2503.20384*, 2025.

[97] Hao Chen, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Renrui Zhang, Xiaoqi Li, Xiao He, Yandong Guo, Chi-Wing Fu, Shanghang Zhang, and Pheng-Ann Heng. Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning. *arXiv preprint arXiv:2506.01953*, 2025.

[98] Hao Li, Shuai Yang, Yilun Chen, Yang Tian, Xiaoda Yang, Xinyi Chen, Hanqing Wang, Tai Wang, Feng Zhao, Dahua Lin, and Jiangmiao Pang. Cronusvla: Transferring latent motion across time for multi-frame prediction in manipulation. *arXiv preprint arXiv:2506.19816*, 2025.

[99] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, Chengkai Hou, Mengdi Zhao, KC alex Zhou, Pheng-Ann Heng, and Shanghang Zhang. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[100] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, Zhizheng Zhang, and He Wang. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

[101] Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025.

[102] Haoming Song, Delin Qu, Yuanqi Yao, Qizhi Chen, Qi Lv, Yiwen Tang, Modi Shi, Guanghui Ren, Maoqing Yao, Bin Zhao, Dong Wang, and Xuelong Li. Hume: Introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*, 2025.

[103] Meng Li, Zhen Zhao, Zhengping Che, Fei Liao, Kun Wu, Zhiyuan Xu, Pei Ren, Zhao Jin, Ning Liu, and Jian Tang. Switchvla: Execution-aware task switching for vision-language-action models. *arXiv preprint arXiv:2506.03574*, 2025.

[104] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[105] Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025.

[106] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

[107] Xiaowei Chi, Kuangzhi Ge, Jiaming Liu, Siyuan Zhou, Peidong Jia, Zichen He, Yuzhen Liu, Tingguang Li, Lei Han, Sirui Han, Shanghang Zhang, and Yike Guo. Mind: Unified visual imagination and control via hierarchical world models. *arXiv preprint arXiv:2506.18897*, 2025.

[108] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 9156–9172. Curran Associates, Inc., 2023.

[109] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 8633–8646. Curran Associates, Inc., 2022.

[110] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

[111] Hongyin Zhang, Pengxiang Ding, Shangke Lyu, Ying Peng, and Donglin Wang. Gevrm: Goal-expressive video generation model for robust visual manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

[112] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 22304–22325. Curran Associates, Inc., 2023.

[113] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 3943–3960. PMLR, 06–09 Nov 2025.

[114] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[115] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning (CoRL)*, Vol. 205 of *Proceedings of Machine Learning Research*, pp. 715–725. PMLR, 14–18 Dec 2023.

[116] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

[117] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023.

[118] Iman Nematollahi, Branton DeMoss, Akshay L Chandra, Nick Hawes, Wolfram Burgard, and Ingmar Posner. Lumos: Language-conditioned imitation learning with world models. *arXiv preprint arXiv:2503.10370*, 2025.

[119] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations (ICLR)*, 2024.

[120] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8121–8130, 2022.

[121] Chuan Wen, Xingyu Lin, John Ian Reyes So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[122] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pp. 18–35, Cham, 2025. Springer Nature Switzerland.

[123] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pp. 306–324, Cham, 2025. Springer Nature Switzerland.

[124] Kanchana Ranasinghe, Xiang Li, Cristina Mata, Jongwoo Park, and Michael S Ryoo. Pixel motion as universal representation for robot control. *arXiv preprint arXiv:2505.07817*, 2025.

[125] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pp. 402–419, Cham, 2020. Springer International Publishing.

[126] Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. *arXiv preprint arXiv:2412.04445*, 2024.

[127] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[128] Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. Uniskill: Imitating human videos via cross-embodiment skill representations. *arXiv preprint arXiv:2505.08787*, 2025.

[129] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

[130] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022.

[131] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[132] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, June 2021.

[133] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 27. Curran Associates, Inc., 2014.

[134] Peiyan Li, Hongtao Wu, Yan Huang, Chilam Cheang, Liang Wang, and Tao Kong. Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 10, No. 2, pp. 1912–1919, 2025.

[135] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025.

[136] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[137] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[138] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[139] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[140] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[141] Changhe Chen, Quantao Yang, Xiaohao Xu, Nima Fazeli, and Olov Andersson. Visa-flow: Accelerating robot skill learning via large-scale video semantic action flow. *arXiv preprint arXiv:2505.01288*, 2025.

[142] James J. Gibson. The theory of affordances. In John Bransford Robert E Shaw, editor, *Perceiving, acting, and knowing: toward an ecological psychology*, pp. pp.67–82. Hillsdale, N.J. : Lawrence Erlbaum Associates, 1977.

[143] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Vol. 229 of *Proceedings of Machine Learning Research*, pp. 540–562. PMLR, 06–09 Nov 2023.

[144] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pp. 728–755, Cham, 2022. Springer Nature Switzerland.

[145] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

[146] Olivia Y. Lee, Annie Xie, Kuan Fang, Karl Pertsch, and Chelsea Finn. Affordance-guided reinforcement learning via visual prompting. *arXiv preprint arXiv:2407.10341*, 2024.

[147] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Vol. 229 of *Proceedings of Machine Learning Research*, pp. 178–200. PMLR, 06–09 Nov 2023.

[148] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pp. 405–421, Cham, 2020. Springer International Publishing.

[149] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

[150] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19729–19739, October 2023.

[151] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[152] Olaolu Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe David Kennedy, and Mac Schwager. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 4748–4770. PMLR, 06–09 Nov 2025.

[153] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, Vol. 42, No. 4, July 2023.

[154] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13778–13790, June 2023.

[155] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The dataset. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pp. 753–771, Cham, 2018. Springer International Publishing.

[156] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, Vol. 130, p. 33–55, 2022.

[157] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[158] Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. Hrp: Human affordances for robotic pre-training. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

[159] Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. *arXiv preprint arXiv:2503.07135*, 2025.

[160] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 4005–4020. PMLR, 06–09 Nov 2025.

[161] Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. Roboground: Robotic manipulation with grounded vision-language priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22540–22550, June 2025.

[162] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024.

[163] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, Yuxuan Kuang, Meng Cao, Feng Zheng, and Xiaodan Liang. A0: An affordance-aware hierarchical model for general robotic manipulation. *arXiv preprint arXiv:2504.12636*, 2025.

[164] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1724–1734, June 2025.

[165] Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Feifei Feng. Improving vision-language-action models via chain-of-affordance. *arXiv preprint arXiv:2412.20451*, 2024.

[166] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[167] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[168] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252, 2015.

[169] Cristoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *Proceedings of Neurips Data-Centric AI Workshop*, 2021.

[170] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 25278–25294. Curran Associates, Inc., 2022.

[171] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

[172] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[173] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 23716–23736. Curran Associates, Inc., 2022.

[174] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025.

[175] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pp. 38–55, Cham, 2025. Springer Nature Switzerland.

[176] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pp. 350–368, Cham, 2022. Springer Nature Switzerland.

[177] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161, June 2024.

[178] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[179] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2020.

[180] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%$*$chatgpt quality, March 2023.

[181] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[182] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[183] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023.

[184] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb

Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

[185] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

[186] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023.

[187] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, June 2025.

[188] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[189] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[190] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[191] Ye Niu, Sanping Zhou, Yizhe Li, Ye Den, and Le Wang. Time-unified diffusion policy with action discrimination for robotic manipulation. *arXiv preprint arXiv:2506.09422*, 2025.

[192] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[193] Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, Cewu Lu, and Wenqiang Zhang. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. *arXiv preprint arXiv:2505.22159*, 2025.

[194] Zexin Zheng, Jia-Feng Cai, Xiao-Ming Wu, Yi-Lin Wei, Yu-Ming Tang, and Wei-Shi Zheng. imanip: Skill-incremental learning for robotic manipulation. *arXiv preprint arXiv:2503.07087*, 2025.

[195] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.

[196] Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22508–22519, June 2025.

[197] Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, and Ziwei Liu. Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters. *arXiv preprint arXiv:2412.00174*, 2024.

[198] Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.

[199] Wei Zhao, Pengxiang Ding, Zhang Min, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. Vlas: Vision-language-action model with speech instructions for customized robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

[200] Renhao Wang, Haoran Geng, Tingle Li, Feishi Wang, Gopala Anumanchipalli, Philipp Wu, Trevor Darrell, Boyi Li, Pieter Abbeel, Jitendra Malik, and Alexei A. Efros. Multigen: Using multimodal generation in simulation to learn multimodal policies in real. *arXiv preprint arXiv:2507.02864*, 2025.

[201] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[202] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proc. Interspeech 2021*, pp. 571–575, 2021.

[203] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023.

[204] Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.

[205] Shunlei Li, Jin Wang, Rui Dai, Wanyu Ma, Wing Yin Ng, Yingbai Hu, and Zheng Li. Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. *arXiv preprint arXiv:2409.19590*, 2024.

[206] Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025.

[207] Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025.

[208] Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

[209] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 5, No. 3, pp. 3838–3845, 2020.

[210] Chaofan Zhang, Shaowei Cui, Shuo Wang, Jingyi Hu, Yinghao Cai, Rui Wang, and Yu Wang. Gelstereo 2.0: An improved gelstereo sensor with multimedium refractive stereo calibration. *IEEE Transactions on Industrial Electronics*, Vol. 71, No. 7, pp. 7452–7462, 2024.

[211] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vol. 235 of *Proceedings of Machine Learning Research*, pp. 14080–14101. PMLR, 21–27 Jul 2024.

[212] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, June 2024.

[213] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[214] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, and Ankit Goyal. Hamster: Hierarchical action models for open-world robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

[215] Wenxuan Song, Jiayi Chen, Wenxue Li, Xu He, Han Zhao, Can Cui, Pengxiang Ding Shiyan Su, Feilong Tang, Xuelian Cheng, Donglin Wang, Zongyuan Ge, Xinhu Zheng, Zhe Liu, Hesheng Wang, and Haoang Li. Rationalvla: A rational vision-language-action model with dual system. *arXiv preprint arXiv:2506.10826*, 2025.

[216] Can Cui, Pengxiang Ding, Wenxuan Song, Shuanghao Bai, Xinyang Tong, Zirui Ge, Runze Suo, Wanqi Zhou, Yang Liu, Bofang Jia, Han Zhao, Siteng Huang, and Donglin Wang. Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation. *arXiv preprint arXiv:2505.03912*, 2025.

[217] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 1949–1974. PMLR, 06–09 Nov 2025.

[218] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025.

[219] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[220] Feng Yan, Fanfan Liu, Liming Zheng, Yufeng Zhong, Yiyang Huang, Zechao Guan, Chengjian Feng, and Lin Ma. Robomm: All-in-one multimodal large model for robotic manipulation. *arXiv preprint arXiv:2412.07215*, 2024.

[221] Haoquan Fang, Markus Grotz, Wilbert Pumacay, Yi Ru Wang, Dieter Fox, Ranjay Krishna, and Jiafei Duan. Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation. *arXiv preprint arXiv:2501.18564*, 2025.

[222] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. RVT-2: Learning Precise Manipulation from Few Demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[223] Ishika Singh, Ankit Goyal, Stan Birchfield, Dieter Fox, Animesh Garg, and Valts Blukis. Og-vla: 3d-aware vision language action model via orthographic image generation. *arXiv preprint arXiv:2506.01196*, 2025.

[224] Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. *arXiv preprint arXiv:2506.07961*, 2025.

[225] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024.

[226] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C. Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025.

[227] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pp. 523–540, Cham, 2020. Springer International Publishing.

[228] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[229] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. Curran Associates, Inc., 2017.

[230] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 23192–23204. Curran Associates, Inc., 2022.

[231] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

[232] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhao Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv preprint arXiv:2502.13143*, 2025.

[233] Yuelei Li, Ge Yan, Annabella Macaluso, Mazeyu Ji, Xueyan Zou, and Xiaolong Wang. Integrating lmm planners and 3d skill policies for generalizable manipulation. *arXiv preprint arXiv:2501.18733*, 2025.

[234] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 1541–1566. PMLR, 06–09 Nov 2025.

[235] Jieyi Zhang, Wenqiang Xu, Zhenjun Yu, Pengfei Xie, Tutian Tang, and Cewu Lu. Dextog: Learning task-oriented dexterous grasp with language. *arXiv preprint arXiv:2504.04573*, 2025.

[236] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, Vol. 7, No. 2, p. 187–199, Apr 2021.

[237] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 31. Curran Associates, Inc., 2018.

[238] Dantong Niu, Yuvan Sharma, Haoru Xue, Giscard Biamby, Junyi Zhang, Ziteng Ji, Trevor Darrell, and Roei Herzig. Pre-training auto-regressive robotic models with 4d representations. *arXiv preprint arXiv:2502.13142*, 2025.

[239] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[240] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 20067–20079. Curran Associates, Inc., 2023.

[241] Dongjiang Li, Bo Peng, Chang Li, Ning Qiao, Qi Zheng, Lei Sun, Yusen Qin, Bangguo Li, Yifeng Luan, Bo Wu, Yibing Zhan, Mingang Sun, Tong Xu, Lusong Li, Hui Shen, and Xiaodong He. An atomic skill library construction method for data-efficient embodied manipulation. *arXiv preprint arXiv:2501.15068*, 2025.

[242] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, Adrian Li-Bell, Danny Driess, Lachy Groom, Sergey Levine, and Chelsea Finn. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[243] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

[244] Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, Yuefan Wang, Huaicheng Zhou, Wenshuo Feng, Jiacheng Liu, Siteng Huang, and Donglin Wang. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

[245] Yi Yang, Jiaxuan Sun, Siqi Kou, Yihan Wang, and Zhijie Deng. Lohovla: A unified vision-language-action model for long-horizon embodied tasks. *arXiv preprint arXiv:2506.00411*, 2025.

[246] ByungOk Han, Jaehong Kim, and Jinhyeok Jang. A dual process vla: Efficient robotic manipulation leveraging vlm. *arXiv preprint arXiv:2410.15549*, 2024.

[247] Zhenyang Liu, Yongchong Gu, Sixiao Zheng, Xiangyang Xue, and Yanwei Fu. Trivla: A triple-system-based unified vision-language-action model for general robot control. *arXiv preprint arXiv:2507.01424*, 2025.

[248] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025.

[249] Zhekai Duan, Yuan Zhang, Shikai Geng, Gaowen Liu, Joschka Boedecker, and Chris Xiaoxuan Lu. Fast ecot: Efficient embodied chain-of-thought via thoughts reuse. *arXiv preprint arXiv:2506.07639*, 2025.

[250] Letian Fu, Huang Huang, Gaurav Datta, Lawrence Yunliang Chen, William Chung-Ho Panitch, Fangchen Liu, Hui Li, and Ken Goldberg. In-context imitation learning via next-token prediction. *arXiv preprint arXiv:2408.15980*, 2024.

[251] Vivek Myers, Bill Chunyuan Zheng, Anca Dragan, Kuan Fang, and Sergey Levine. Temporal representation alignment: Successor features enable emergent compositionality in robot instruction following. *arXiv preprint arXiv:2502.05454*, 2025.

[252] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[253] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018.

[254] Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. *arXiv preprint arXiv:2501.16664*, 2025.

[255] Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *arXiv preprint arXiv:2502.05450*, 2025.

[256] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. Serl: A software suite for sample-efficient robotic reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16961–16969, 2024.

[257] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2410.21845*, 2024.

[258] Weiqiao Han, Sergey Levine, and Pieter Abbeel. Learning compound multi-step controllers under unknown dynamics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6435–6442, 2015.

[259] Abhishek Gupta, Justin Yu, Tony Z. Zhao, Vikash Kumar, Aaron Rovinsky, Kelvin Xu, Thomas Devlin, and Sergey Levine. Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6664–6671, 2021.

[260] Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 1577–1594. PMLR, 23–29 Jul 2023.

[261] Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025.

[262] Charles Xu, Qiyang Li, Jianlan Luo, and Sergey Levine. Rldg: Robotic generalist policy distillation via reinforcement learning. *arXiv preprint arXiv:2412.09858*, 2024.

[263] Yuxuan Chen and Xiao Li. Rlrc: Reinforcement learning-based recovery for compressed vision-language-action models. *arXiv preprint arXiv:2506.17639*, 2025.

[264] Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv preprint arXiv:2506.15799*, 2025.

[265] Haichao Zhang, Haonan Yu, Le Zhao, Andrew Choi, Qinxun Bai, Break Yang, and Wei Xu. Slim: Sim-to-real legged instructive manipulation via long-horizon visuomotor learning. *arXiv preprint arXiv:2501.09905*, 2025.

[266] Tobias Jülg, Wolfram Burgard, and Florian Walter. Refined policy distillation: From vla generalists to rl experts. *arXiv preprint arXiv:2503.05833*, 2025.

[267] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing.

[268] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[269] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[270] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei

Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

[271] Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U-Xuan Tan, Navonil Majumder, and Soujanya Poria. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

[272] Cunxin Fan, Xiaosong Jia, Yihang Sun, Yixiao Wang, Jianglan Wei, Ziyang Gong, Xiangyu Zhao, Masayoshi Tomizuka, Xue Yang, Junchi Yan, and Mingyu Ding. Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions. *arXiv preprint arXiv:2505.02152*, 2025.

[273] Peng Chen, Pi Bu, Yingyao Wang, Xinyi Wang, Ziming Wang, Jie Guo, Yingxiu Zhao, Qi Zhu, Jun Song, Siran Yang, Jiamang Wang, and Bo Zheng. Combatvla: An efficient vision-language-action model for combat tasks in 3d action role-playing games. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[274] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 34892–34916. Curran Associates, Inc., 2023.

[275] Wei Zhao, Gongsheng Li, Zhefei Gong, Pengxiang Ding, Han Zhao, and Donglin Wang. Unveiling the potential of vision-language-action models with open-ended multimodal instructions. *arXiv preprint arXiv:2505.11214*, 2025.

[276] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, December 2024. [Online; accessed 2025-08-04].

[277] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[278] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023.

[279] Pengxiang Ding, Han Zhao, Wenjie Zhang, Wenxuan Song, Min Zhang, Siteng Huang, Ningxi Yang, and Donglin Wang. Quar-vla: Vision-language-action model for quadruped robots. *arXiv preprint arXiv:2312.14457*, 2023.

[280] Han Zhao, Wenxuan Song, Donglin Wang, Xinyang Tong, Pengxiang Ding, Xuelian Cheng, and Zongyuan Ge. More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. *arXiv preprint arXiv:2503.08007*, 2025.

[281] Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37, pp. 56619–56643. Curran Associates, Inc., 2024.

[282] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun

Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Moham-

mad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[283] Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. *arXiv preprint arXiv:2502.01828*, 2025.

[284] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

[285] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations (ICLR)*, 2025.

[286] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025.

[287] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, June 2025.

[288] Oleg Sautenkov, Yasheerah Yaqoot, Artem Lykov, Muhammad Ahsan Mustafa, Grik Tadevosyan, Aibek Akhmetkazy, Miguel Altamirano Cabrera, Mikhail Martynov, Sausar Karaf, and Dzmitry Tsetserukou. Uav-vla: Vision-language-action system for large scale aerial mission generation. *arXiv preprint arXiv:2501.05014*, 2025.

[289] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26689–26699, June 2024.

[290] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2025.

[291] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki,

Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.

[292] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2025.

[293] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025.

[294] Sombit Dey, Jan-Nico Zaech, Nikolay Nikolov, Luc Van Gool, and Danda Pani Paudel. Revla: Reverting visual domain limitation of robotic foundation models. *arXiv preprint arXiv:2409.15250*, 2024.

[295] Joey Hejna, Chethan Anand Bhateja, Yichen Jiang, Karl Pertsch, and Dorsa Sadigh. Remix: Optimizing data mixtures for large scale imitation learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 145–164. PMLR, 06–09 Nov 2025.

[296] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[297] Hongyu Wang, Chuyan Xiong, Ruiping Wang, and Xilin Chen. Bitvla: 1-bit vision-language-action models for robotics manipulation. *arXiv preprint arXiv:2506.07530*, 2025.

[298] Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *arXiv preprint arXiv:2506.07339*, 2025.

[299] Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *arXiv preprint arXiv:2411.02359*, 2024.

[300] Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation. *arXiv preprint arXiv:2502.02175*, 2025.

[301] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[302] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 4066–4083. PMLR, 06–09 Nov 2025.

[303] ALOHA 2 Team, Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sanky Chan, Kenneth Draper, Debidatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, Wayne Gramlich, Torr Hage, Alexander Herzog, Jonathan Hoech, Thinh Nguyen, Ian Storz, Baruch Tabanpour, Leila Takayama, Jonathan Tompson, Ayzaan Wahid, Ted Wahrburg, Sichun Xu, Sergey Yaroshenko, Kevin Zakka, and Tony Z. Zhao. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024.

[304] Tony Z. Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Seyed Kamyar Seyed Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 1910–1924. PMLR, 06–09 Nov 2025.

[305] Thanpimon Buamanee, Masato Kobayashi, Yuki Uranishi, and Haruo Takemura. Bi-act: Bilateral control-based imitation learning via action chunking with transformer. In *2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 410–415, 2024.

[306] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12156–12163, 2024.

[307] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[308] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[309] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, Nathan Ratliff, and Dieter Fox. Curobo: Parallelized collision-free robot motion generation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8112–8119, 2023.

[310] Shiqi Yang, Minghuan Liu, Yuzhe Qin, Runyu Ding, Jialong Li, Xuxin Cheng, Ruihan Yang, Sha Yi, and Xiaolong Wang. Ace: A cross-platform and visual-exoskeletons system for low-cost dexterous teleoperation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 4895–4911. PMLR, 06–09 Nov 2025.

[311] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 2729–2749. PMLR, 06–09 Nov 2025.

[312] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. *arXiv preprint arXiv:2407.03162*, 2024.

[313] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[314] TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, Naveen Kuppuswamy, Kuan-Hui Lee, Katherine Liu, Dale McConachie, Ian McMahon, Haruki Nishimura, Calder Phillips-Grafflin, Charles Richter, Paarth Shah, Krishnan Srinivasan, Blake Wulfe, Chen Xu, Mengchao Zhang, Alex Alspach, Maya Angeles, Kushal Arora, Vitor Campagnolo Guizilini, Alejandro Castro, Dian Chen, Ting-Sheng Chu, Sam Creasey, Sean Curtis, Richard Denitto, Emma Dixon, Eric Dusel, Matthew Ferreira, Aimee Goncalves, Grant Gould, Damrong Guoy, Swati Gupta, Xuchen Han, Kyle Hatch, Brendan Hathaway, Allison Henry, Hillel Hochsztein, Phoebe Horgan, Shun Iwase, Donovon Jackson, Siddharth Karamcheti, Sedrick Keh, Joseph Masterjohn, Jean Mercat, Patrick Miller, Paul Mitiguy, Tony Nguyen, Jeremy Nimmer, Yuki Noguchi, Reko Ong, Aykut Onol, Owen Pfannenstiehl, Richard Poyner, Leticia Priebe Mendes Rocha, Gordon Richardson, Christopher Rodriguez, Derick Seale, Michael Sherman, Mariah Smith-Jones, David Tago, Pavel Tokmakov, Matthew Tran, Basile Van Hoorick, Igor Vasiljevic, Sergey Zakharov, Mark Zolotas, Rares Ambrus, Kerri Fetzer-Borelli, Benjamin Burchfiel, Hadas Kress-Gazit, Siyuan Feng, Stacie Ford, and Russ Tedrake. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

[315] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. In *3rd RSS Workshop on Dexterous Manipulation: Learning and Control with Diverse Data*, 2025.

[316] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[317] Haritheja Etukuru, Norihito Naka, Zijin Hu, Seungjae Lee, Julian Mehu, Aaron Edsinger, Chris Paxton, Soumith Chintala, Lerrel Pinto, and Nur Muhammad Mahi Shafiullah. Robot utility models: General policies for zero-shot deployment in new environments. *arXiv preprint arXiv:2409.05865*, 2024.

[318] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[319] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *arXiv preprint arXiv:2505.07813*, 2025.

[320] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja,

Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

[321] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071, June 2025.

[322] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[323] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

[324] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.

[325] Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. Egozero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*, 2025.

[326] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[327] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, Vol. 87 of *Proceedings of Machine Learning Research*, pp. 879–893. PMLR, 29–31 Oct 2018.

[328] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Anol Bhattacherjee and Brian Fitzgerald, editors, *Shaping the Future of ICT Research. Methods and Approaches*, pp. 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[329] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters (RA-L)*, pp. 1–8, 2023.

[330] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen,

Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[331] Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, and Soujanya Poria. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14199–14214, Vienna, Austria, July 2025. Association for Computational Linguistics.

[332] Nils Blank, Moritz Reuss, Marcel Rühle, Ömer Erdinç Yağmurlu, Fabian Wenzel, Oier Mees, and Rudolf Lioutikov. Scaling robot policy learning via zero-shot labeling with foundation models. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 4158–4187. PMLR, 06–09 Nov 2025.

[333] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, Shichao Fan, Xinhua Wang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Ning Liu, Pei Ren, Qiang Zhang, Yaoxu Lyu, Mengzhen Liu, Jingyang He, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2025.

[334] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano,

HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush

Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[335] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[336] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, Vol. 87 of *Proceedings of Machine Learning Research*, pp. 651–673. PMLR, 29–31 Oct 2018.

[337] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[338] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning (CoRL)*, Vol. 100 of *Proceedings of Machine Learning Research*, pp. 885–897. PMLR, 30 Oct–01 Nov 2020.

[339] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, New York City, NY, USA, June 2022.

[340] Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Vol. 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 06–09 Nov 2023.

[341] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning (CoRL)*, Vol. 164 of *Proceedings of Machine Learning Research*, pp. 991–1002. PMLR, 08–11 Nov 2022.

[342] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 7, No. 3, pp. 7327–7334, 2022.

[343] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 44776–44791. Curran Associates, Inc., 2023.

[344] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21013–21022, June 2022.

[345] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456, June 2024.

[346] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10138–10148, October 2021.

[347] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[348] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[349] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[350] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

[351] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2022.

[352] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.

[353] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Vol. 229 of *Proceedings of Machine Learning Research*, pp. 1820–1864. PMLR, 06–09 Nov 2023.

[354] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2025.

[355] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, Vol. 87 of *Proceedings of Machine Learning Research*, pp. 906–915. PMLR, 29–31 Oct 2018.

[356] Sabela Ramos, Sertan Girgin, Léonard Hussenot, Damien Vincent, Hanna Yakubovich, Daniel Toyama, Anita Gergely, Piotr Stanczyk, Raphael Marinier, Jeremiah Harmsen, Olivier Pietquin, and Nikola Momchev. Rlds: an ecosystem to generate, share and use datasets in reinforcement learning. *arXiv preprint arXiv:2111.02767*, 2021.

[357] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning (CoRL)*, Vol. 205 of *Proceedings of Machine Learning Research*, pp. 1838–1849. PMLR, 14–18 Dec 2023.

[358] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11576–11582, 2023.

[359] Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023.

[360] Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multistage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, Vol. 40, pp. 1476–1491, 2024.

[361] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. `https://sites.google.com/view/berkeley-ur5/home`.

[362] Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, Chelsea Finn, and Abhinav Gupta. Train offline, test online: A real robot learning benchmark. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9197–9203, 2023.

[363] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4788–4795, 2024.

[364] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 9, No. 1, pp. 49–56, 2024.

[365] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 7, No. 4, pp. 11807–11814, 2022.

[366] Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning (CoRL)*, Vol. 164 of *Proceedings of Machine Learning Research*, pp. 674–684. PMLR, 08–11 Nov 2022.

[367] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[368] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652, 2024.

[369] Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. Cacti: A framework for scalable multi-task multi-scene visual imitation learning. In *CoRL 2022 Workshop on Pre-training Robot Learning*, 2022.

[370] Qiuyu Chen, Shosuke C Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[371] Tianhe Yu, Ted Xiao, Jonathan Tompson, Austin Stone, Su Wang, Anthony Brohan, Jaspiar Singh, Clayton Tan, Dee M, Jodilyn Peralta, Karol Hausman, Brian Ichter, and Fei Xia. Scaling robot learning with semantically imagined experience. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[372] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18359–18369, June 2023.

[373] Asher J. Hancock, Allen Z. Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. *arXiv preprint arXiv:2410.01971*, 2024.

[374] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Robotic Skill Acquisition via Instruction Augmentation with Vision-Language Models. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[375] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Vol. 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[376] Liyiming Ke, Yunchu Zhang, Abhay Deshpande, Siddhartha Srinivasa, and Abhishek Gupta. Ccil: Continuity-based data augmentation for corrective imitation learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[377] Maged Iskandar, Christian Ott, Oliver Eiberger, Manuel Keppler, Alin Albu-Schäffer, and Alexander Dietrich. Joint-level control of the dlr lightweight robot sara. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8903–8910, 2020.

[378] Simon Guist, Jan Schneider, Hao Ma, Le Chen, Vincent Berenz, Julian Martus, Heiko Ott, Felix Grüninger, Michael Muehlebach, Jonathan Fiene, Bernhard Schölkopf, and Dieter Büchler. Safe & Accurate at Speed with Tendons: A Robot Arm for Exploring Dynamic Motion. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[379] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. LEAP Hand: Low-Cost, Efficient, and Anthropomorphic Hand for Robot Learning. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[380] Muhamamd Haris Khan, Selamawit Asfaw, Dmitrii Iarchuk, Miguel Altamirano Cabrera, Luis Moreno, Issatay Tokmurziyev, and Dzmitry Tsetserukou. Shake-vla: Vision-language-action model-based system for bimanual robotic manipulations and liquid mixing. *arXiv preprint arXiv:2501.06919*, 2025.

[381] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, Yifeng Zhu, and Kevin Lin. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

[382] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning (CoRL)*, Vol. 164 of *Proceedings of Machine Learning Research*, pp. 1678–1690. PMLR, 08–11 Nov 2022.

[383] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-Scale Simulation of Household Tasks for Generalist Robots. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[384] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning (CoRL)*, Vol. 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100. PMLR, 30 Oct–01 Nov 2020.

[385] Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Screenath, and Shankar Sastry. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.

[386] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1, 2021.

[387] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations (ICLR)*, 2023.

[388] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

[389] Arth Shukla, Stone Tao, and Hao Su. Maniskill-hab: A benchmark for low-level manipulation in home rearrangement tasks. In *International Conference on Learning Representations (ICLR)*, 2025.

[390] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27649–27660, June 2025.

[391] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[392] Shengqiang Zhang, Philipp Wicke, Lütfi Kerem Şenel, Luis Figueredo, Abdeldjallil Naceri, Sami Haddadin, Barbara Plank, and Hinrich Schütze. Lohoravens: A long-horizon language-conditioned benchmark for robotic tabletop manipulation. *arXiv preprint arXiv:2310.12020*, 2023.

[393] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[394] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In

M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34, pp. 251–266. Curran Associates, Inc., 2021.

[395] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[396] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 5, No. 2, pp. 3019–3026, 2020.

[397] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[398] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv*, 2017.

[399] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, Ranjay Krishna, Dustin Schwenk, Eli VanderBilt, and Aniruddha Kembhavi. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16238–16250, June 2024.

[400] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning (CoRL)*, Vol. 270 of *Proceedings of Machine Learning Research*, pp. 3705–3728. PMLR, 06–09 Nov 2025.

[401] Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, Jonathan Tremblay, Kanav Arora, Kirsty Ellis, Luca Macesanu, Matthew Leonard, Meedeum Cho, Ozgur Aslan, Shivin Dass, Jie Wang, Xingfang Yuan, Xuning Yang, Abhishek Gupta, Dinesh Jayaraman, Glen Berseth, Kostas Daniilidis, Roberto Martin-Martin, Youngwoon Lee, Percy Liang, Chelsea Finn, and Sergey Levine. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025.

[402] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

[403] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters (RA-L)*, Vol. 8, No. 6, pp. 3740–3747, 2023.

[404] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[405] Liming Zheng, Feng Yan, Fanfan Liu, Chengjian Feng, Zhuoliang Kang, and Lin Ma. Robocas: A benchmark for robotic manipulation in complex object arrangement scenarios. In *NeurIPS: Datasets and Benchmarks Track*, 2024.

[406] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21190–21200, 2023.

[407] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016.

[408] Eric Rohmer, Surya P. N. Singh, and Marc Freese. V-rep: A versatile and scalable robot simulation framework. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1321–1326, 2013.

[409] Stephen James, Marc Freese, and Andrew J. Davison. Pyrep: Bringing v-rep to deep robot learning. *arXiv preprint arXiv:1906.11176*, 2019.

[410] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. Robothor: An open simulation-to-real embodied ai platform. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[411] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 5982–5994. Curran Associates, Inc., 2022.

[412] Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. Vlatest: Testing and evaluating vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12894*, 2024.

[413] Taowen Wang, Cheng Han, James Chenhao Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. *arXiv preprint arXiv:2411.13587*, 2024.

[414] Hao Cheng, Erjia Xiao, Chengyuan Yu, Zhao Yao, Jiahang Cao, Qiang Zhang, Jiaxu Wang, Mengshu Sun, Kaidi Xu, Jindong Gu, and Renjing Xu. Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models. *arXiv preprint arXiv:2409.13174*, 2024.

[415] Hong Lu, Hengxu Li, Prithviraj Singh Shahani, Stephanie Herbers, and Matthias Scheutz. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. *arXiv preprint arXiv:2502.04558*, 2025.

[416] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, Fei Xia, Jasmine Hsu, Jonathan Hoech, Pete Florence, Sean Kirmani, Sumeet Singh, Vikas Sindhwani, Carolina Parada, Chelsea Finn, Peng Xu, Sergey Levine, and Jie Tan. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024.

[417] Valerii Serpiva, Artem Lykov, Artyom Myshlyaev, Muhammad Haris Khan, Ali Alridha Abdulkarim, Oleg Sautenkov, and Dzmitry Tsetserukou. Racevla: Vla-based racing drone navigation with human-like behaviour. *arXiv preprint arXiv:2503.02572*, 2025.

[418] Artem Lykov, Valerii Serpiva, Muhammad Haris Khan, Oleg Sautenkov, Artyom Myshlyaev, Grik Tadevosyan, Yasheerah Yaqoot, and Dzmitry Tsetserukou. Cognitivedrone: A vla model and evaluation benchmark for real-time cognitive task solving and reasoning in uavs. *arXiv preprint arXiv:2503.01378*, 2025.

[419] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

[420] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

[421] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

[422] Frederik Nolte, Bernhard Schölkopf, and Ingmar Posner. Is single-view mesh reconstruction ready for robotics? *arXiv preprint arXiv:2505.17966*, 2025.

[423] Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. Safe: Multitask failure detection for vision-language-action models. *arXiv preprint arXiv:2506.09937*, 2025.

[424] Zhejian Yang, Yongchao Chen, Xueyang Zhou, Jiangyue Yan, Dingjie Song, Yinuo Liu, Yuting Li, Yu Zhang, Pan Zhou, Hechang Chen, and Lichao Sun. Agentic robot: A brain-inspired framework for vision-language-action models in embodied agents. *arXiv preprint arXiv:2505.23450*, 2025.

[425] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pp. 1–7, 2025.