Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications

Kento Kawaharazuka
The University of Tokyo
kawaharazuka@jsk.imi.i.u-tokyo.ac.jp

Jihoon Oh
The University of Tokyo
oh@jsk.imi.i.u-tokyo.ac.jp

Jun Yamada
University of Oxford
jyamada@robots.ox.ac.uk

Ingmar Posner*
University of Oxford
ingmar@robots.ox.ac.uk

Yuke Zhu*
The University of Texas at Austin
yukez@cs.utexas.edu

***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.**

***Index Terms*—Vision-Language-Action Models, Robotics, Foundation Models, Imitation Learning, Robot Learning**

I. 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.

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. I.1). While

*Equal advising

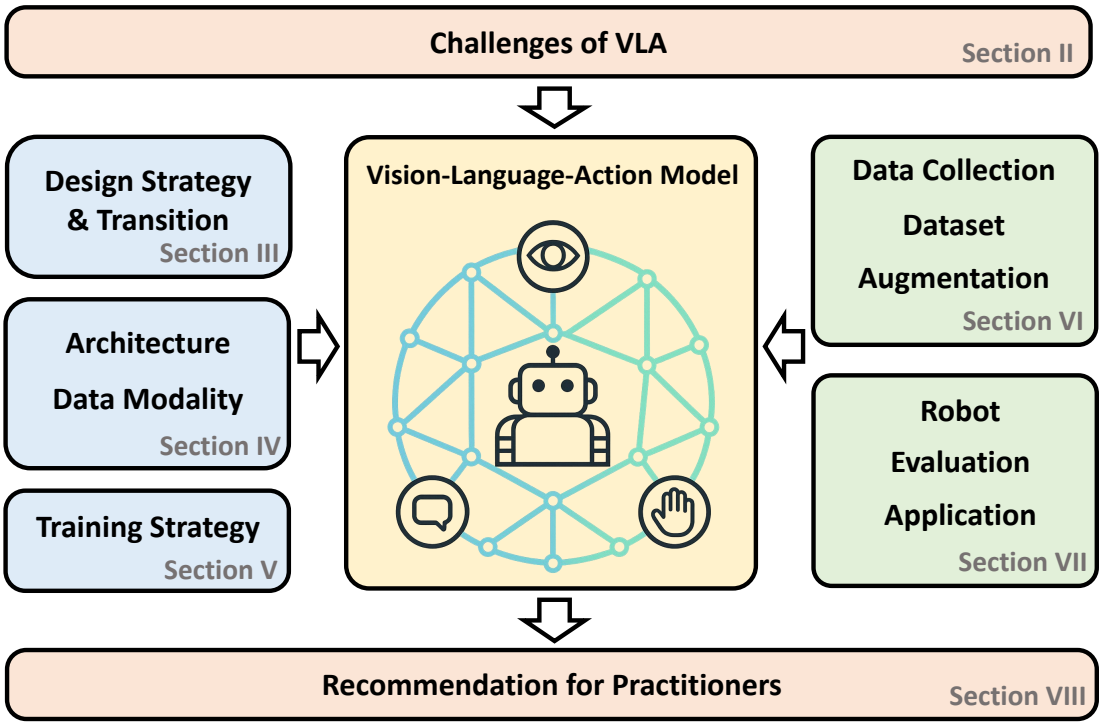

Fig. 1. **Structure of this survey.** Section II outlines the key challenges in developing Vision-Language-Action (VLA) models. Section III and Section IV review the evolution of VLA strategies, architectures, and modality-specific design choices. Section V categorizes training strategies and practical implementation considerations. Section VI discusses the data collection methologies, publicly available dataset, and data augmentation. Section VII 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 VIII, based on the findings of the systematic review.

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 I.1 (Vision-Language-Action (VLA) Model)**.** ***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 II outlines the key challenges addressed in VLA research. Section III reviews major strategies and the architectural transition of VLA models. Section IV introduces core architectural components and building blocks, including modality-specific processing modules. Section V discusses key training strategies and practical implementation considerations. Section VI summarises data collection methodologies, publicly available datasets, and data augmentation. Then, Section VII 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 VIII. Finally, Section IX discusses open challenges and future directions, and Section X presents our concluding remarks.

II. 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.

A. 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.

B. 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?

C. 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.

III. 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-

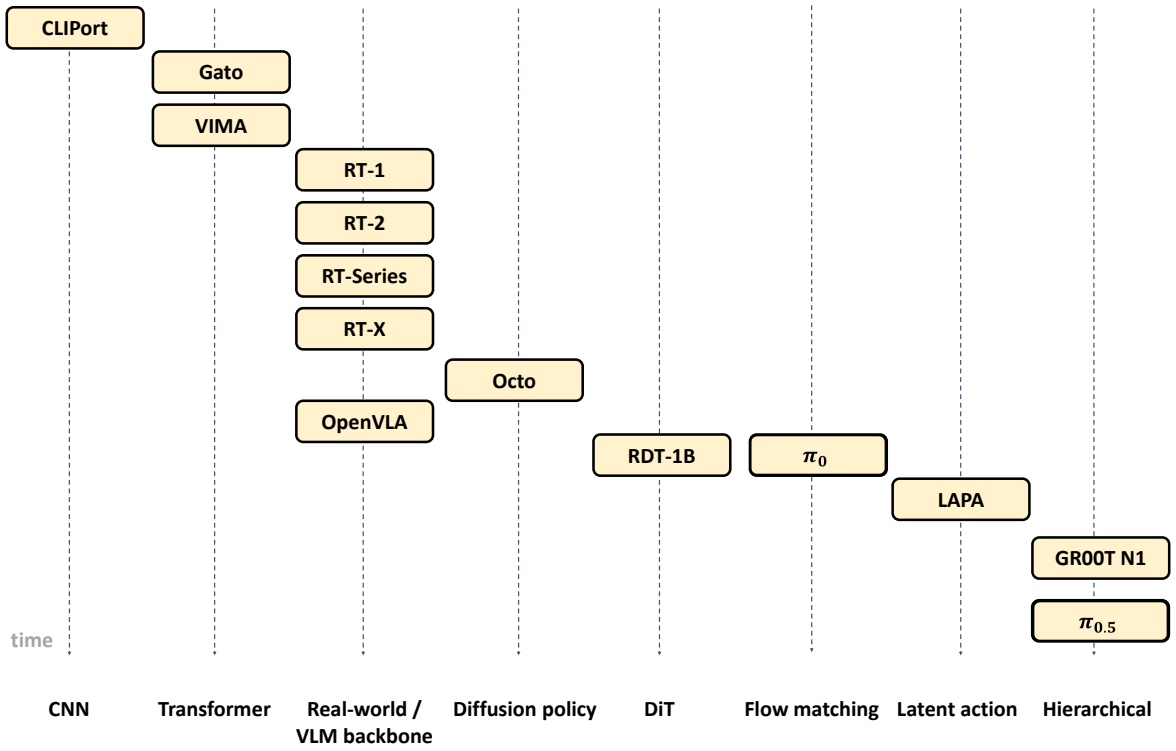

Fig. 2. **Timeline of major Vision–Language–Action (VLA) models.** This figure summarizes the historical progression of representative VLA models shown in Section III: from early CNN-based models (e.g., *CLIPort* [15]), to real-world scalable policies leveraging pre-trained VLM backbones (e.g., *RT-1*, *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]).

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 IV-A). 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 computes their difference. VQ-VAE [53] is applied to this difference, generating a discrete token $\boldsymbol{z}_t$ which, together with $\boldsymbol{x}_t$, is used to reconstruct $\boldsymbol{x}_{t+H}$. 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 $\boldsymbol{z}_t$. 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.

IV. 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.

A. Sensorimotor Model

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

(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

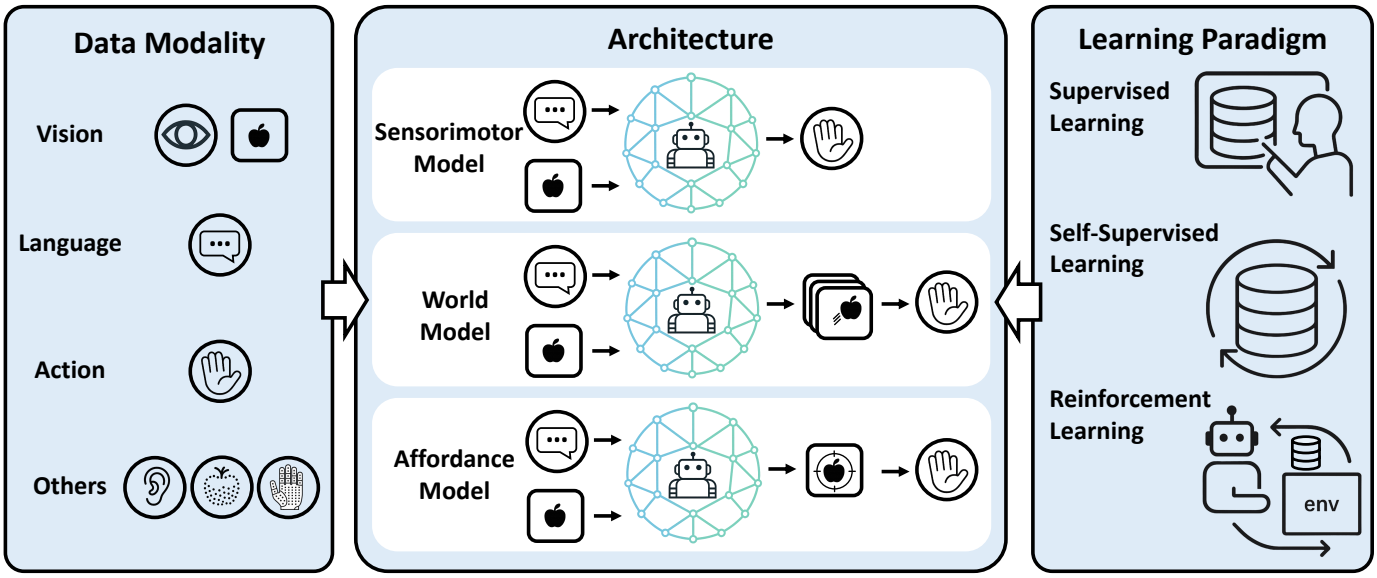

Fig. 3. **Structure of Section IV and Section V.** The figure summarizes key components of VLA models. The center illustrates core architectural types, including sensorimotor models, world models, and affordance-based models. The left side depicts the input and output modalities—vision, language, action,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

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

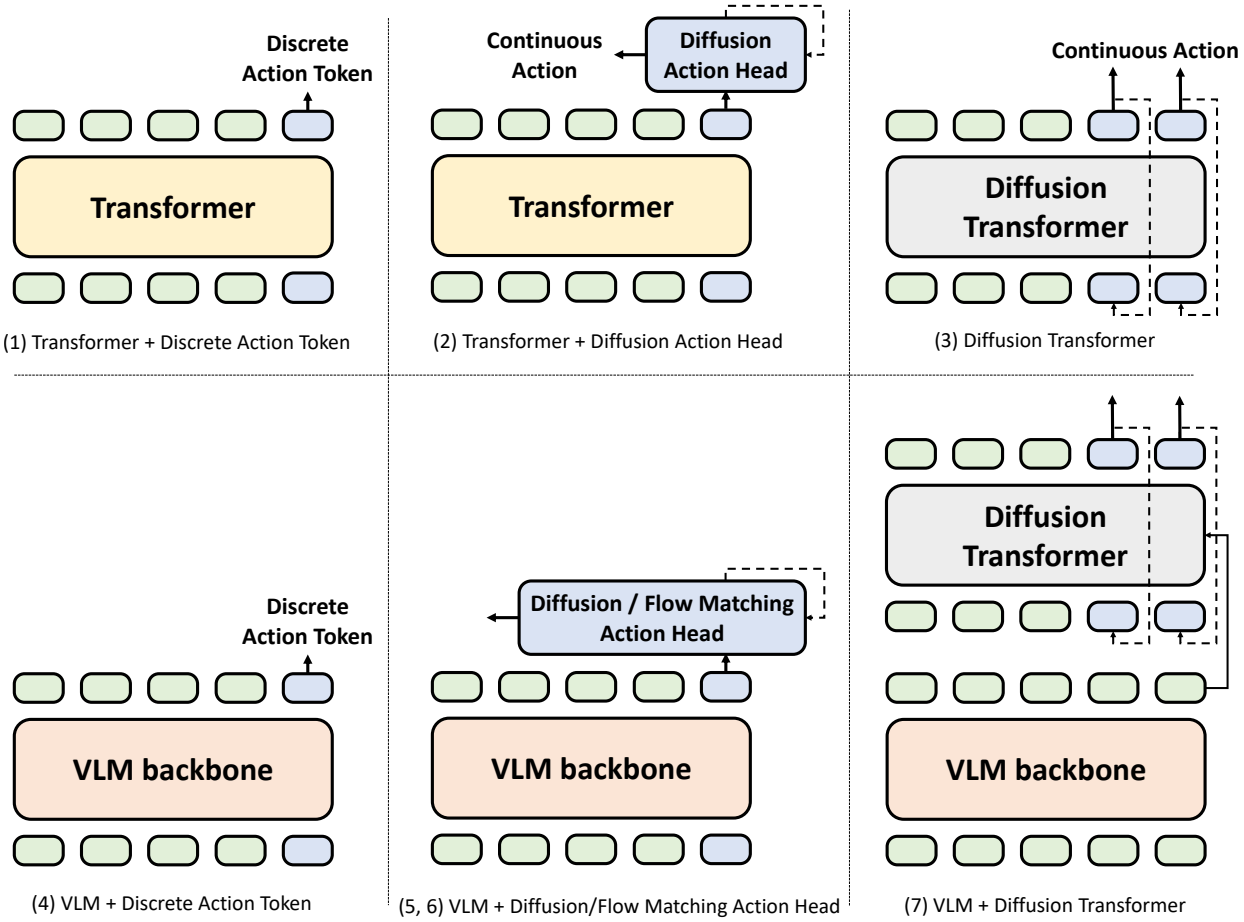

Fig. 4. **Architecture of sensorimotor models for VLA.** This figure categorizes seven representative architectures used in recent VLA research. (1) *Transformer + Discrete Action Token*: A standard transformer processes tokenized inputs to predict discrete actions. (2) *Transformer + Diffusion Action 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.

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].

B. 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] es-

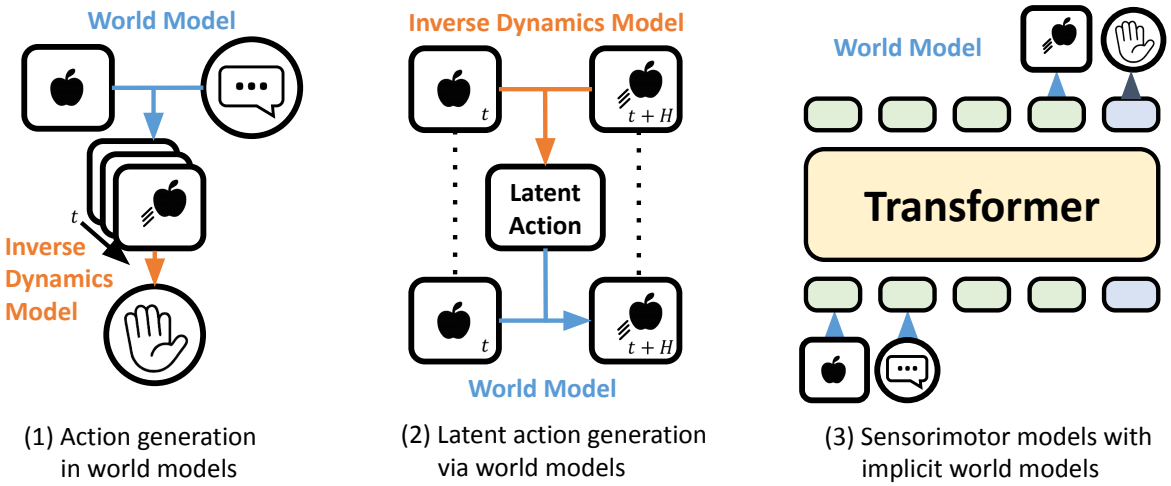

Fig. 5. **Design patterns for incorporating world models in VLA.** (1) Using world models in conjunction with inverse dynamics models to generate actions. (2) Leveraging world models to learn latent action representations, particularly from human videos; the resulting latent tokens are then used for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

timates 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 IV-A 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.

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 III 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 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].

C. 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 IV-A), 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.

D. 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.

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,

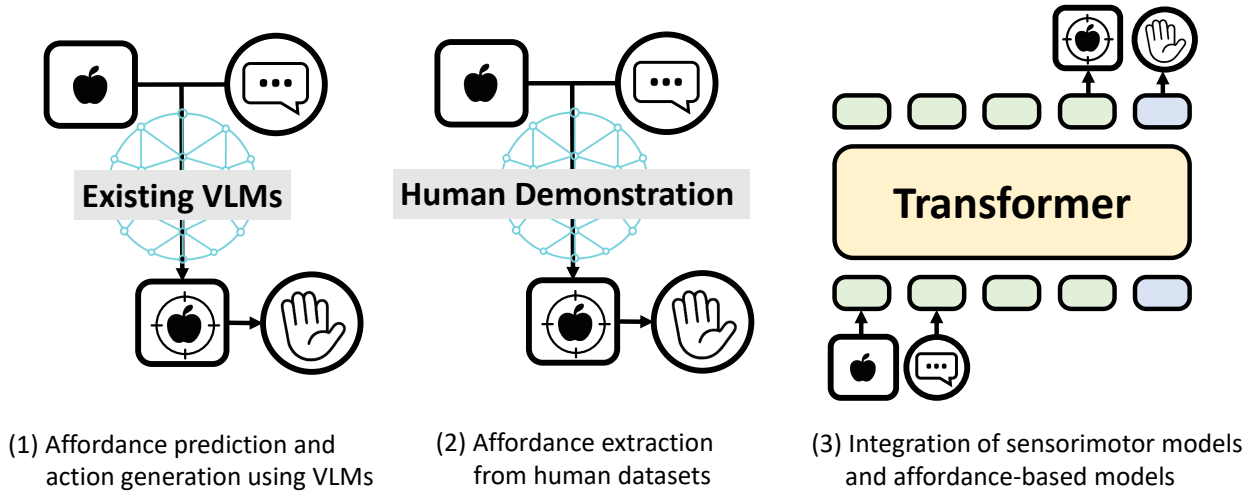

Fig. 6. **Design patterns for incorporating affordance-based models in VLA.** (1) Predicting affordances and subsequently generating actions conditioned on the predicted affordances; (2) Extracting affordances from human demonstration videos and learning latent representations to guide action generation;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

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].

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].

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 IV-B.

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) 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 IV-D1) 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: Struct-Diffusion [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 SplatMOVER [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.

E. 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 Humanoid-VLA [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.

V. 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.

A. 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 V-D1.

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.

B. 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

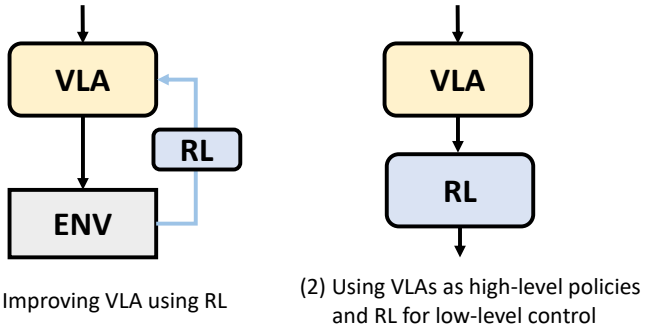

Fig. 7. **Approaches to integrating RL with VLA models.** (1) RL is used to fine-tune VLA models to enhance their performance. (2) VLA models serve as high-level policies, while RL policies handle low-level control.

(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 IV-B and Section IV-D3. 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.

C. 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 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.

D. 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.

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 Hybrid-VLA [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.

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. 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 pre-trained representations, potentially losing valuable general knowledge that could benefit zero-shot generalization.

E. 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.

VI. Datasets

A. 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 VI-B; 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

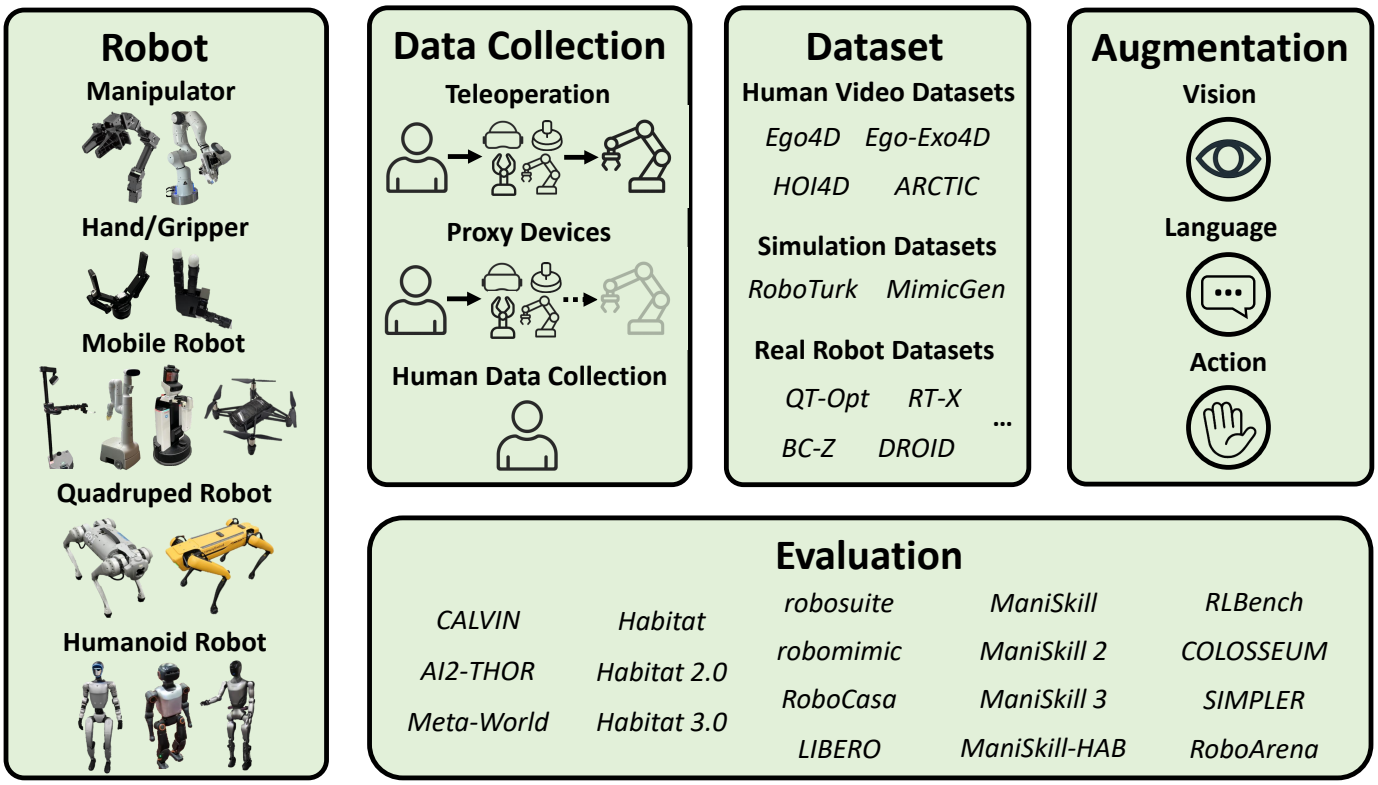

Fig. 8. **Structure of Section VI and Section VII:** robots used VLA research — including manipulator, hand/gripper, mobile robot, quadruped robot, and humanoid robot; data collection methods — including teleoperation, proxy devices, and human data collection; publicly available dataset — including human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

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 demonstrations 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.

B. 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 VI-A). 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],

TABLE I

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

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.

C. 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.

VII. 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.

A. 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.

B. 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.

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

TABLE II

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

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. LoHo-Ravens [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.

C. 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 Cognitive-Drone [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 Track-VLA [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.

VIII. 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.

IX. Future Research Direction

A. 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.

B. 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.

C. 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.

D. 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.

E. 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.

F. 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.

G. 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.

H. 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.

X. 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.

ACKNOWLEDGEMENTS

This research was partially supported by JST CRONOS under Grant Number JPMJCS24K6. ChatGPT-4o was used to assist in the creation of several figures.

REFERENCES

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2024.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 19 730–19 742.

[4] OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[5] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, H.-S. Fang *et al.*, "Toward general-purpose robots via foundation models: A survey and meta-analysis," 2023.

[6] K. Kawaharazuka, T. Matsushima, A. Gambardella, J. G. Guo, C. Paxton, and A. Zeng, "Real-world robot applications of foundation models: A review," *Advanced Robotics*, vol. 38, no. 18, pp. 1232–1254, 2024.

[7] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. 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] b. ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Proceedings of The 6th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 287–318.

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

[10] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 2165–2183.

[11] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[12] R. Sapkota, Y. Cao, K. I. Roumeliotis, and M. Karkee, "Vision-language-action models: Concepts, progress, applications and challenges," *arXiv preprint arXiv:2505.04769*, 2025.

[13] Y. Zhong, F. Bai, S. Cai, X. Huang, Z. Chen, X. Zhang, Y. Wang, S. Guo, T. Guan, K. N. Lui *et al.*, "A survey on vision-language-action models: An action tokenization perspective," *arXiv preprint arXiv:2507.01925*, 2025.

[14] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[15] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 894–906. [Online]. Available: https://proceedings.mlr.press/v164/shridhar22a.html

[16] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "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] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903.

[18] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[19] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo *et al.*, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

[20] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," in *International Conference on Learning Representations (ICLR)*, 2025.

[21] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[22] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin *et al.*, "Latent action pretraining from videos," in *International Conference on Learning Representations (ICLR)*, 2025.

[23] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[24] J. Bjorck, F. C. neda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[26] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Proceedings of the 2020 Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 726–747.

[27] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg *et al.*, "A generalist agent," *Transactions on Machine Learning Research*, 2022.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[29] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds. Brussels, Belgium: Association for Computational Linguistics, November 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012/

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[31] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "VIMA: Robot manipulation

with multimodal prompts," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 14 975–15 022. [Online]. Available: https://proceedings.mlr.press/v202/jiang23b.html

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. 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. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[35] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

[36] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[37] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: Adaptive space-time tokenization for videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 786–12 797.

[38] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 8469–8488.

[39] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay *et al.*, "On scaling up a multilingual vision and language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 432–14 444.

[40] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman *et al.*, "Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 70–96.

[41] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu *et al.*, "Rt-trajectory: Robotic task generalization via hindsight trajectory sketches," in *International Conference on Learning Representations (ICLR)*, 2024.

[42] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh, "Rt-h: Action hierarchies using language," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[43] I. Leal, K. Choromanski, D. Jain, A. Dubey, J. Varley, M. Ryoo, Y. Lu, F. Liu, V. Sindhwani, Q. Vuong *et al.*, "Sara-rt: Scaling up robotics transformers with self-adaptive robust attention," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6920–6927.

[44] M. Ahn, D. Dwibedi, C. Finn, M. G. Arenas, K. Gopalakrishnan, K. Hausman, B. Ichter, A. Irpan, N. Joshi, R. Julian *et al.*, "Autort: Embodied foundation models for large scale orchestration of robotic agents," *arXiv preprint arXiv:2401.12963*, 2024.

[45] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic VLMs: Investigating the design space of visually-conditioned language models," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 23 123–23 144. [Online]. Available: https://proceedings.mlr.press/v235/karamcheti24a.html

[46] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2024.

[47] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 975–11 986.

[48] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. C. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

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

[50] C. Zhou, L. YU, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model," in *International Conference on Learning Representations (ICLR)*, 2025.

[51] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[52] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations (ICLR)*, 2023.

[53] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[54] H. Liu, W. Yan, M. Zaharia, and P. Abbeel, "World model on million-length video and language with ringattention," *arXiv preprint*, 2024.

[55] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

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

[57] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia *et al.*, "Open-world object manipulation using pre-trained vision-language models," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 3397–3417.

[58] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman *et al.*, "Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches," *arXiv preprint arXiv:2403.02709*, 2024.

[59] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu *et al.*, "Rt-trajectory: Robotic task generalization via hindsight trajectory sketches," *arXiv preprint arXiv:2311.01977*, 2023.

[60] K. Bousmalis, G. Vezzani, D. Rao, C. M. Devin, A. X. Lee, M. B. Villalonga, T. Davchev, Y. Zhou, A. Gupta, A. Raju *et al.*, "Robocat: A self-improving generalist agent for robotic manipulation," *Transactions on Machine Learning Research*, 2024.

[61] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu *et al.*, "Vision-language foundation models as effective robot imitators," *arXiv preprint arXiv:2311.01378*, 2023.

[62] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. 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)*, June 2024, pp. 26 439–26 455.

[63] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. 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] S. Haldar, Z. Peng, and L. Pinto, "Baku: An efficient transformer for multi-task policy learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Globerson, L. Mackey, D. Belgrave,

A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 141 208–141 239.

[65] Y. Ma, D. Chi, S. Wu, Y. Liu, Y. Zhuang, J. Hao, and I. King, "Actra: Optimized transformer architecture for vision-language-action models in robot learning," *arXiv preprint arXiv:2408.01147*, 2024.

[66] X. Pang, W. Xia, Z. Wang, B. Zhao, D. Hu, D. Wang, and X. 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)*, 2024, pp. 7251–7256.

[67] R. Doshi, H. R. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 496–512.

[68] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[69] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 63–70.

[70] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen *et al.*, "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] S. Wang, S. Liu, W. Wang, J. Shan, and B. Fang, "Robobert: An end-to-end multimodal robotic manipulation model," *arXiv preprint arXiv:2502.07837*, 2025.

[72] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. 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)*, June 2025, pp. 27 661–27 672.

[73] W. Liu, Y. Du, T. Hermans, S. Chernova, and C. 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] M. Reuss, Ömer Erdinç Yağmurlu, F. Wenzel, and R. Lioutikov, "Multimodal Diffusion Transformer: Learning Versatile Behavior from Multimodal Goals," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[75] Y. Zhong, X. Huang, R. Li, C. Zhang, Y. Liang, Y. Yang, and Y. Chen, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *arXiv preprint arXiv:2502.20900*, 2025.

[76] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025.

[77] R. Yang, G. Chen, C. Wen, and Y. Gao, "Fp3: A 3d foundation policy for robotic manipulation," *arXiv preprint arXiv:2503.08950*, 2025.

[78] Y. Yao, S. Liu, H. Song, D. Qu, Q. Chen, Y. Ding, B. Zhao, Z. Wang, X. Li, and D. 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)*, June 2025, pp. 22 573–22 583.

[79] Y. Yang, Z. Cai, Y. Tian, J. Zeng, and J. Pang, "Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation," *arXiv preprint arXiv:2504.17784*, 2025.

[80] Z. Hou, T. Zhang, Y. Xiong, H. Pu, C. Zhao, R. Tong, Y. Qiao, J. Dai, and Y. Chen, "Diffusion transformer policy," *arXiv preprint arXiv:2410.15959*, 2024.

[81] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An embodied generalist agent in 3D world," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 20 413–20 451. [Online]. Available: https://proceedings.mlr.press/v235/huang24ae.html

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

[83] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 40 085–40 110.

[84] P. Ding, H. Zhao, W. Zhang, W. Song, M. Zhang, S. Huang, N. Yang, and D. Wang, "Quar-vla: Vision-language-action model for quadruped robots," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 352–367.

[85] X. Li, C. Mata, J. Park, K. Kahatapitiya, Y. S. Jang, J. Shang, K. Ranasinghe, R. Burgert, M. Cai, Y. J. Lee *et al.*, "Llara: Supercharging robot learning data for vision-language policy," in *International Conference on Learning Representations (ICLR)*, 2025.

[86] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 3157–3181.

[87] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.

[88] F. Liu, F. Yan, L. Zheng, C. Feng, Y. Huang, and L. Ma, "Robouniview: Visual-language model with unified view representation for robotic manipulation," *arXiv preprint arXiv:2406.18977*, 2024.

[89] H. Arai, K. Miwa, K. Sasaki, K. Watanabe, Y. Yamaguchi, S. Aoki, and I. Yamamoto, "Covla: Comprehensive vision-language-action dataset for autonomous driving," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, February 2025, pp. 1933–1943.

[90] J. Wen, M. Zhu, Y. Zhu, Z. Tang, J. Li, Z. Zhou, C. Li, X. Liu, Y. Peng, C. Shen *et al.*, "Diffusion-vla: Generalizable and interpretable robot foundation model via self-generated reasoning," *arXiv preprint arXiv:2412.03293*, 2024.

[91] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv preprint arXiv:2502.05855*, 2025.

[92] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, R. Cheng, Y. Peng, C. Shen *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," *arXiv preprint arXiv:2502.14420*, 2025.

[93] M. Zhu, Y. Zhu, J. Li, Z. Zhou, J. Wen, X. Liu, C. Shen, Y. Peng, and F. Feng, "Objectvla: End-to-end open-world object manipulation without demonstration," *arXiv preprint arXiv:2502.19250*, 2025.

[94] AgiBot-World-Contributors, Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[95] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," *arXiv preprint arXiv:2503.07511*, 2025.

[96] R. Zhang, M. Dong, Y. Zhang, L. Heng, X. Chi, G. Dai, L. Du, Y. Du, and S. 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] H. Chen, J. Liu, C. Gu, Z. Liu, R. Zhang, X. Li, X. He, Y. Guo, C.-W. Fu, S. Zhang *et al.*, "Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning," *arXiv preprint arXiv:2506.01953*, 2025.

[98] H. Li, S. Yang, Y. Chen, Y. Tian, X. Yang, X. Chen, H. Wang, T. Wang, F. Zhao, D. Lin *et al.*, "Cronusvla: Transferring latent motion across time for multi-frame prediction in manipulation," *arXiv preprint arXiv:2506.19816*, 2025.

[99] J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, X. Li, Z. Guo, S. Chen, M. Liu *et al.*, "Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model," *arXiv preprint arXiv:2503.10631*, 2025.

[100] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui *et al.*, "Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv preprint arXiv:2505.03233*, 2025.

[101] F. Lin, R. Nai, Y. Hu, J. You, J. Zhao, and Y. Gao, "Onetwovla: A unified vision-language-action model with adaptive reasoning," *arXiv preprint arXiv:2505.11917*, 2025.

[102] H. Song, D. Qu, Y. Yao, Q. Chen, Q. Lv, Y. Tang, M. Shi, G. Ren, M. Yao, B. Zhao *et al.*, "Hume: Introducing system-2 thinking in visual-language-action model," *arXiv preprint arXiv:2505.21432*, 2025.

[103] M. Li, Z. Zhao, Z. Che, F. Liao, K. Wu, Z. Xu, P. Ren, Z. Jin, N. Liu, and J. Tang, "Switchvla: Execution-aware task switching for vision-language-action models," *arXiv preprint arXiv:2506.03574*, 2025.

[104] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv preprint arXiv:2411.19650*, 2024.

[105] S. Wang, J. Zhang, M. Li, J. Liu, A. Li, K. Wu, F. Zhong, J. Yu, Z. Zhang, and H. Wang, "Trackvla: Embodied visual tracking in the wild," *arXiv preprint arXiv:2505.23189*, 2025.

[106] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[107] X. Chi, K. Ge, J. Liu, S. Zhou, P. Jia, Z. He, Y. Liu, T. Li, L. Han, S. Han *et al.*, "Mind: Unified visual imagination and control via hierarchical world models," *arXiv preprint arXiv:2506.18897*, 2025.

[108] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 9156–9172.

[109] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 8633–8646.

[110] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin *et al.*, "Dreamgen: Unlocking generalization in robot learning through video world models," *arXiv preprint arXiv:2505.12705*, 2025.

[111] H. Zhang, P. Ding, S. Lyu, Y. Peng, and D. Wang, "Gevrm: Goal-expressive video generation model for robust visual manipulation," in *International Conference on Learning Representations (ICLR)*, 2025.

[112] A. Ajay, S. Han, Y. Du, S. Li, A. Gupta, T. Jaakkola, J. Tenenbaum, L. Kaelbling, A. Srivastava, and P. Agrawal, "Compositional foundation models for hierarchical planning," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 22 304–22 325.

[113] J. Liang, R. Liu, E. Ozguroglu, S. Sudhakar, A. Dave, P. Tokmakov, S. Song, and C. Vondrick, "Dreamitate: Real-world visuomotor policy learning via video generation," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 3943–3960.

[114] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[115] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," in *Proceedings of The 6th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 715–725. [Online]. Available: https://proceedings.mlr.press/v205/labbe23a.html

[116] K. Black, M. Nakamoto, P. Atreya, H. R. Walke, C. Finn, A. Kumar, and S. Levine, "Zero-shot robotic manipulation with pre-trained image-editing diffusion models," in *International Conference on Learning Representations (ICLR)*, 2024.

[117] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 392–18 402.

[118] I. Nematollahi, B. DeMoss, A. L. Chandra, N. Hawes, W. Burgard, and I. Posner, "Lumos: Language-conditioned imitation learning with world models," *arXiv preprint arXiv:2503.10370*, 2025.

[119] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," in *International Conference on Learning Representations (ICLR)*, 2024.

[120] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8121–8130.

[121] C. Wen, X. Lin, J. I. R. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[122] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It is better to track together," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 18–35.

[123] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 306–324.

[124] K. Ranasinghe, X. Li, C. Mata, J. Park, and M. S. Ryoo, "Pixel motion as universal representation for robot control," *arXiv preprint arXiv:2505.07817*, 2025.

[125] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 402–419.

[126] Y. Chen, Y. Ge, W. Tang, Y. Li, Y. Ge, M. Ding, Y. Shan, and X. Liu, "Moto: Latent motion token as the bridging language for learning robot manipulation from videos," *arXiv preprint arXiv:2412.04445*, 2024.

[127] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "Univla: Learning to act anywhere with task-centric latent actions," *arXiv preprint arXiv:2505.06111*, 2025.

[128] H. Kim, J. Kang, H. Kang, M. Cho, S. J. Kim, and Y. Lee, "Uniskill: Imitating human videos via cross-embodiment skill representations," *arXiv preprint arXiv:2505.08787*, 2025.

[129] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 000–16 009.

[130] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "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)*, June 2022, pp. 18 995–19 012.

[131] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[132] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 873–12 883.

[133] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems (NeurIPS)*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[134] P. Li, H. Wu, Y. Huang, C. Cheang, L. Wang, and T. 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] C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma *et al.*, "Gr-3 technical report," *arXiv preprint arXiv:2507.15493*, 2025.

[136] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[137] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[138] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[139] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, F. Hu, J. Jang, K. Kundalia, Z. Lin, L. Magne *et al.*, "Flare: Robot learning with implicit world modeling," *arXiv preprint arXiv:2505.15659*, 2025.

[140] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang *et al.*, "Worldvla: Towards autoregressive action world model," *arXiv preprint arXiv:2506.21539*, 2025.

[141] C. Chen, Q. Yang, X. Xu, N. Fazeli, and O. Andersson, "Visa-flow: Accelerating robot skill learning via large-scale video semantic action flow," *arXiv preprint arXiv:2505.01288*, 2025.

[142] J. J. Gibson, "The theory of affordances," in *Perceiving, acting, and knowing: toward an ecological psychology*, J. B. Robert E Shaw, Ed. Hillsdale, N.J. : Lawrence Erlbaum Associates, 1977, pp. pp.67–82. [Online]. Available: https://hal.science/hal-00692033

[143] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 540–562.

[144] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 728–755.

[145] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[146] O. Y. Lee, A. Xie, K. Fang, K. Pertsch, and C. Finn, "Affordance-guided reinforcement learning via visual prompting," *arXiv preprint arXiv:2407.10341*, 2024.

[147] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 178–200. [Online]. Available: https://proceedings.mlr.press/v229/rashid23a.html

[148] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 405–421.

[149] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9650–9660.

[150] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 729–19 739.

[151] H.-S. Fang, C. Wang, M. Gou, and C. 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] O. Shorinwa, J. Tucker, A. Smith, A. Swann, T. Chen, R. Firoozi, M. D. Kennedy, and M. Schwager, "Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 4748–4770. [Online]. Available: https://proceedings.mlr.press/v270/shorinwa25a.html

[153] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[154] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. 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)*, June 2023, pp. 13 778–13 790.

[155] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The dataset," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 753–771.

[156] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision (IJCV)*, vol. 130, p. 33–55, 2022. [Online]. Available: https://doi.org/10.1007/s11263-021-01531-2

[157] D. Shan, J. Geng, M. Shu, and D. 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] M. K. Srirama, S. Dasari, S. Bahl, and A. Gupta, "Hrp: Human affordances for robotic pre-training," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

[159] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. 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] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, "Robopoint: A vision-language model for spatial affordance prediction in robotics," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 4005–4020.

[161] H. Huang, X. Chen, Y. Chen, H. Li, X. Han, Z. Wang, T. Wang, J. Pang, and Z. Zhao, "Roboground: Robotic manipulation with grounded vision-language priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 22 540–22 550.

[162] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao, "Rt-affordance: Affordances are versatile intermediate representations for robot manipulation," *arXiv preprint arXiv:2411.02704*, 2024.

[163] R. Xu, J. Zhang, M. Guo, Y. Wen, H. Yang, M. Lin, J. Huang, Z. Li, K. Zhang, L. Wang *et al.*, "A0: An affordance-aware hierarchical model for general robotic manipulation," *arXiv preprint arXiv:2504.12636*, 2025.

[164] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An *et al.*, "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)*, June 2025, pp. 1724–1734.

[165] J. Li, Y. Zhu, Z. Tang, J. Wen, M. Zhu, X. Liu, C. Li, R. Cheng, Y. Peng, and F. Feng, "Improving vision-language-action models via chain-of-affordance," *arXiv preprint arXiv:2412.20451*, 2024.

[166] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[167] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[168] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[169] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," in *Proceedings of Neurips Data-Centric AI Workshop*, 2021.

[170] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 25 278–25 294.

[171] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2818–2829.

[172] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[173] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23 716–23 736.

[174] H. Fu, D. Zhang, Z. Zhao, J. Cui, D. Liang, C. Zhang, D. Zhang, H. Xie, B. Wang, and X. Bai, "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation," *arXiv preprint arXiv:2503.19755*, 2025.

[175] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 38–55.

[176] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 350–368.

[177] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing, "Putting the object back into video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 3151–3161.

[178] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, November 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

[179] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2020.

[180] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%∗chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[181] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[182] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[183] M. Javaheripi, S. Bubeck, M. Abdin, J. Aneja, S. Bubeck, C. C. T. Mendes, W. Chen, A. Del Giorno, R. Eldan, S. Gopi *et al.*, "Phi-2: The surprising power of small language models," *Microsoft Research Blog*, 2023.

[184] L. B. Allal, A. Lozhkov, E. Bakouch, G. M. Blázquez, G. Penedo, L. Tunstall, A. Marafioti, H. Kydlíček, A. P. Lajarín, V. Srivastav *et al.*, "Smollm2: When smol goes big – data-centric training of a small language model," *arXiv preprint arXiv:2502.02737*, 2025.

[185] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang *et al.*, "Gpt-neox-20b: An open-source autoregressive language model," in *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022. [Online]. Available: https://arxiv.org/abs/2204.06745

[186] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 2397–2430. [Online]. Available: https://proceedings.mlr.press/v202/biderman23a.html

[187] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn *et al.*, "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)*, June 2025, pp. 1702–1713.

[188] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

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

[190] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021.

[191] Y. Niu, S. Zhou, Y. Li, Y. Den, and L. Wang, "Time-unified diffusion policy with action discrimination for robotic manipulation," *arXiv preprint arXiv:2506.09422*, 2025.

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

[193] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai *et al.*, "Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation," *arXiv preprint arXiv:2505.22159*, 2025.

[194] Z. Zheng, J.-F. Cai, X.-M. Wu, Y.-L. Wei, Y.-M. Tang, and W.-S. Zheng, "imanip: Skill-incremental learning for robotic manipulation," *arXiv preprint arXiv:2503.07087*, 2025.

[195] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," *arXiv preprint arXiv:2402.19432*, 2024.

[196] J. Zheng, J. Li, D. Liu, Y. Zheng, Z. Wang, Z. Ou, Y. Liu, J. Liu, Y.-Q. Zhang, and X. Zhan, "Universal actions for enhanced embodied foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 22 508–22 519.

[197] J. Jiang, W. Xiao, Z. Lin, H. Zhang, T. Ren, Y. Gao, Z. Lin, Z. Cai, L. Yang, and Z. Liu, "Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters," *arXiv preprint arXiv:2412.00174*, 2024.

[198] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," *arXiv preprint arXiv:2501.04693*, 2025.

[199] W. Zhao, P. Ding, Z. Min, Z. Gong, S. Bai, H. Zhao, and D. Wang, "Vlas: Vision-language-action model with speech instructions for customized robot manipulation," in *International Conference on Learning Representations (ICLR)*, 2025.

[200] R. Wang, H. Geng, T. Li, F. Wang, G. Anumanchipalli, P. Wu, T. Darrell, B. Li, P. Abbeel, J. Malik *et al.*, "Multigen: Using multimodal generation in simulation to learn multimodal policies in real," *arXiv preprint arXiv:2507.02864*, 2025.

[201] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[202] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[203] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[204] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[205] S. Li, J. Wang, R. Dai, W. Ma, W. Y. Ng, Y. Hu, and Z. Li, "Robonurse-vla: Robotic scrub nurse system based on vision-language-action model," *arXiv preprint arXiv:2409.19590*, 2024.

[206] P. Hao, C. Zhang, D. Li, X. Cao, X. Hao, S. Cui, and S. Wang, "Tla: Tactile-language-action model for contact-rich manipulation," *arXiv preprint arXiv:2503.08548*, 2025.

[207] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv preprint arXiv:2505.09577*, 2025.

[208] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, "Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv preprint arXiv:2507.09160*, 2025.

[209] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "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] C. Zhang, S. Cui, S. Wang, J. Hu, Y. Cai, R. Wang, and Y. 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] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Goldberg, "A touch, vision, and language dataset for multimodal alignment," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 14 080–14 101. [Online]. Available: https://proceedings.mlr.press/v235/fu24b.html

[212] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. 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)*, June 2024, pp. 10 371–10 381.

[213] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[214] Y. Li, Y. Deng, J. Zhang, J. Jang, M. Memmel, C. R. Garrett, F. Ramos, D. Fox, A. Li, A. Gupta *et al.*, "Hamster: Hierarchical action models for open-world robot manipulation," in *International Conference on Learning Representations (ICLR)*, 2025.

[215] W. Song, J. Chen, W. Li, X. He, H. Zhao, C. Cui, P. D. S. Su, F. Tang, X. Cheng, D. Wang *et al.*, "Rationalvla: A rational vision-language-action model with dual system," *arXiv preprint arXiv:2506.10826*, 2025.

[216] C. Cui, P. Ding, W. Song, S. Bai, X. Tong, Z. Ge, R. Suo, W. Zhou, Y. Liu, B. Jia *et al.*, "Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation," *arXiv preprint arXiv:2505.03912*, 2025.

[217] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 1949–1974.

[218] T. Lin, G. Li, Y. Zhong, Y. Zou, and B. Zhao, "Evo-0: Vision-language-action model with implicit spatial understanding," *arXiv preprint arXiv:2507.00416*, 2025.

[219] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[220] F. Yan, F. Liu, L. Zheng, Y. Zhong, Y. Huang, Z. Guan, C. Feng, and L. Ma, "Robomm: All-in-one multimodal large model for robotic manipulation," *arXiv preprint arXiv:2412.07215*, 2024.

[221] H. Fang, M. Grotz, W. Pumacay, Y. R. Wang, D. Fox, R. Krishna, and J. Duan, "Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation," *arXiv preprint arXiv:2501.18564*, 2025.

[222] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y.-W. Chao, and D. Fox, "RVT-2: Learning Precise Manipulation from Few Demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[223] I. Singh, A. Goyal, S. Birchfield, D. Fox, A. Garg, and V. Blukis, "Og-vla: 3d-aware vision language action model via orthographic image generation," *arXiv preprint arXiv:2506.01196*, 2025.

[224] P. Li, Y. Chen, H. Wu, X. Ma, X. Wu, Y. Huang, L. Wang, T. Kong, and T. Tan, "Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models," *arXiv preprint arXiv:2506.07961*, 2025.

[225] J. Wei, S. Yuan, P. Li, Q. Hu, Z. Gan, and W. Ding, "Occllama: An occupancy-language-action generative world model for autonomous driving," *arXiv preprint arXiv:2409.03272*, 2024.

[226] X. Zhou, X. Han, F. Yang, Y. Ma, and A. C. Knoll, "Opendrivevla: Towards end-to-end autonomous driving with large vision language action model," *arXiv preprint arXiv:2503.23463*, 2025.

[227] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 523–540.

[228] C. R. Qi, H. Su, K. Mo, and L. 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] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NeurIPS)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[230] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23 192–23 204.

[231] J. Zhou, J. Wang, B. Ma, Y.-S. Liu, T. Huang, and X. Wang, "Uni3d: Exploring unified 3d representation at scale," in *International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=wcaE4Dfgt8

[232] Z. Qi, W. Zhang, Y. Ding, R. Dong, X. Yu, J. Li, L. Xu, B. Li, X. He, G. Fan *et al.*, "Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation," *arXiv preprint arXiv:2502.13143*, 2025.

[233] Y. Li, G. Yan, A. Macaluso, M. Ji, X. Zou, and X. Wang, "Integrating lmm planners and 3d skill policies for generalizable manipulation," *arXiv preprint arXiv:2501.18733*, 2025.

[234] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 1541–1566.

[235] J. Zhang, W. Xu, Z. Yu, P. Xie, T. Tang, and C. Lu, "Dextog: Learning task-oriented dexterous grasp with language," *arXiv preprint arXiv:2504.04573*, 2025.

[236] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, p. 187–199, Apr 2021. [Online]. Available: http://dx.doi.org/10.1007/s41095-021-0229-5

[237] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[238] D. Niu, Y. Sharma, H. Xue, G. Biamby, J. Zhang, Z. Ji, T. Darrell, and R. Herzig, "Pre-training auto-regressive robotic models with 4d representations," *arXiv preprint arXiv:2502.13142*, 2025.

[239] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. 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] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motiongpt: Human motion as a foreign language," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 20 067–20 079.

[241] D. Li, B. Peng, C. Li, N. Qiao, Q. Zheng, L. Sun, Y. Qin, B. Li, Y. Luan, B. Wu *et al.*, "An atomic skill library construction method for data-efficient embodied manipulation," *arXiv preprint arXiv:2501.15068*, 2025.

[242] L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai *et al.*, "Hi robot: Open-ended instruc-

tion following with hierarchical vision-language-action models," *arXiv preprint arXiv:2502.19417*, 2025.
[243] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," *arXiv preprint arXiv:2412.04453*, 2024.
[244] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu *et al.*, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv preprint arXiv:2502.14795*, 2025.
[245] Y. Yang, J. Sun, S. Kou, Y. Wang, and Z. Deng, "Lohovla: A unified vision-language-action model for long-horizon embodied tasks," *arXiv preprint arXiv:2506.00411*, 2025.
[246] B. Han, J. Kim, and J. Jang, "A dual process vla: Efficient robotic manipulation leveraging vlm," *arXiv preprint arXiv:2410.15549*, 2024.
[247] Z. Liu, Y. Gu, S. Zheng, X. Xue, and Y. Fu, "Trivla: A triple-system-based unified vision-language-action model for general robot control," *arXiv preprint arXiv:2507.01424*, 2025.
[248] W. Chen, S. Belkhale, S. Mirchandani, O. Mees, D. Driess, K. Pertsch, and S. Levine, "Training strategies for efficient embodied reasoning," *arXiv preprint arXiv:2505.08243*, 2025.
[249] Z. Duan, Y. Zhang, S. Geng, G. Liu, J. Boedecker, and C. X. Lu, "Fast ecot: Efficient embodied chain-of-thought via thoughts reuse," *arXiv preprint arXiv:2506.07639*, 2025.
[250] L. Fu, H. Huang, G. Datta, L. Y. Chen, W. C.-H. Panitch, F. Liu, H. Li, and K. Goldberg, "In-context imitation learning via next-token prediction," *arXiv preprint arXiv:2408.15980*, 2024.
[251] V. Myers, B. C. Zheng, A. Dragan, K. Fang, and S. Levine, "Temporal representation alignment: Successor features enable emergent compositionality in robot instruction following," *arXiv preprint arXiv:2502.05454*, 2025.
[252] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[253] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870.
[254] Y. Guo, J. Zhang, X. Chen, X. Ji, Y.-J. Wang, Y. Hu, and J. Chen, "Improving vision-language-action model with online reinforcement learning," *arXiv preprint arXiv:2501.16664*, 2025.
[255] Y. Chen, S. Tian, S. Liu, Y. Zhou, H. Li, and D. Zhao, "Conrft: A reinforced fine-tuning method for vla models via consistency policy," *arXiv preprint arXiv:2502.05450*, 2025.
[256] J. Luo, Z. Hu, C. Xu, Y. L. Tan, J. Berg, A. Sharma, S. Schaal, C. Finn, A. Gupta, and S. Levine, "Serl: A software suite for sample-efficient robotic reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 961–16 969.
[257] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *arXiv preprint arXiv:2410.21845*, 2024.
[258] W. Han, S. Levine, and P. Abbeel, "Learning compound multi-step controllers under unknown dynamics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 6435–6442.
[259] A. Gupta, J. Yu, T. Z. Zhao, V. Kumar, A. Rovinsky, K. Xu, T. Devlin, and S. 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)*, 2021, pp. 6664–6671.
[260] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, "Efficient online reinforcement learning with offline data," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 1577–1594.
[261] G. Lu, W. Guo, C. Zhang, Y. Zhou, H. Jiang, Z. Gao, Y. Tang, and Z. Wang, "Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning," *arXiv preprint arXiv:2505.18719*, 2025.
[262] C. Xu, Q. Li, J. Luo, and S. Levine, "Rldg: Robotic generalist policy distillation via reinforcement learning," *arXiv preprint arXiv:2412.09858*, 2024.
[263] Y. Chen and X. Li, "Rlrc: Reinforcement learning-based recovery for compressed vision-language-action models," *arXiv preprint arXiv:2506.17639*, 2025.
[264] A. Wagenmaker, M. Nakamoto, Y. Zhang, S. Park, W. Yagoub, A. Nagabandi, A. Gupta, and S. Levine, "Steering your diffusion policy with latent space reinforcement learning," *arXiv preprint arXiv:2506.15799*, 2025.
[265] H. Zhang, H. Yu, L. Zhao, A. Choi, Q. Bai, B. Yang, and W. Xu, "Slim: Sim-to-real legged instructive manipulation via long-horizon visuomotor learning," *arXiv preprint arXiv:2501.09905*, 2025.
[266] T. Jülg, W. Burgard, and F. Walter, "Refined policy distillation: From vla generalists to rl experts," *arXiv preprint arXiv:2503.05833*, 2025.
[267] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.
[268] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
[269] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding *et al.*, "Cosmos world foundation model platform for physical ai," *arXiv preprint arXiv:2501.03575*, 2025.
[270] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2025.
[271] C.-Y. Hung, Q. Sun, P. Hong, A. Zadeh, C. Li, U.-X. Tan, N. Majumder, and S. Poria, "Nora: A small open-sourced generalist vision language action model for embodied tasks," *arXiv preprint arXiv:2504.19854*, 2025.
[272] C. Fan, X. Jia, Y. Sun, Y. Wang, J. Wei, Z. Gong, X. Zhao, M. Tomizuka, X. Yang, J. Yan *et al.*, "Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions," *arXiv preprint arXiv:2505.02152*, 2025.
[273] P. Chen, P. Bu, Y. Wang, X. Wang, Z. Wang, J. Guo, Y. Zhao, Q. Zhu, J. Song, S. Yang *et al.*, "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] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916.
[275] W. Zhao, G. Li, Z. Gong, P. Ding, H. Zhao, and D. Wang, "Unveiling the potential of vision-language-action models with open-ended multimodal instructions," *arXiv preprint arXiv:2505.11214*, 2025.
[276] G. DeepMind. (2024, Dec.) Introducing gemini 2.0: our new ai model for the agentic era. [Online; accessed 2025-08-04]. [Online]. Available: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/
[277] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv preprint arXiv:2503.20020*, 2025.
[278] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar, "Introducing our multimodal models," 2023. [Online]. Available: https://www.adept.ai/blog/fuyu-8b
[279] P. Ding, H. Zhao, W. Zhang, W. Song, M. Zhang, S. Huang, N. Yang, and D. Wang, "Quar-vla: Vision-language-action model for quadruped robots," *arXiv preprint arXiv:2312.14457*, 2023.
[280] H. Zhao, W. Song, D. Wang, X. Tong, P. Ding, X. Cheng, and Z. Ge, "More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models," *arXiv preprint arXiv:2503.08007*, 2025.
[281] Y. Yue, Y. Wang, B. Kang, Y. Han, S. Wang, S. Song, J. Feng, and G. Huang, "Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 56 619–56 643.
[282] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[283] Y. Wu, R. Tian, G. Swamy, and A. Bajcsy, "From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment," *arXiv preprint arXiv:2502.01828*, 2025.

[284] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," *arXiv preprint arXiv:2402.12226*, 2024.

[285] R. Zheng, Y. Liang, S. Huang, J. Gao, H. D. III, A. Kolobov, F. Huang, and J. Yang, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," in *International Conference on Learning Representations (ICLR)*, 2025.

[286] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "Up-vla: A unified understanding and prediction model for embodied agent," *arXiv preprint arXiv:2501.18867*, 2025.

[287] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini *et al.*, "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)*, June 2025, pp. 91–104.

[288] O. Sautenkov, Y. Yaqoot, A. Lykov, M. A. Mustafa, G. Tadevosyan, A. Akhmetkazy, M. A. Cabrera, M. Martynov, S. Karaf, and D. Tsetserukou, "Uav-vla: Vision-language-action system for large scale aerial mission generation," *arXiv preprint arXiv:2501.05014*, 2025.

[289] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 689–26 699.

[290] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2025.

[291] Z. Li, G. Chen, S. Liu, S. Wang, V. VS, Y. Ji, S. Lan, H. Zhang, Y. Zhao, S. Radhakrishnan *et al.*, "Eagle 2: Building post-training data strategies from scratch for frontier vision-language models," *arXiv preprint arXiv:2501.14818*, 2025.

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

[293] D. Driess, J. T. Springenberg, B. Ichter, L. Yu, A. Li-Bell, K. Pertsch, A. Z. Ren, H. Walke, Q. Vuong, L. X. Shi *et al.*, "Knowledge insulating vision-language-action models: Train fast, run fast, generalize better," *arXiv preprint arXiv:2505.23705*, 2025.

[294] S. Dey, J.-N. Zaech, N. Nikolov, L. V. Gool, and D. P. Paudel, "Revla: Reverting visual domain limitation of robotic foundation models," *arXiv preprint arXiv:2409.15250*, 2024.

[295] J. Hejna, C. A. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh, "Remix: Optimizing data mixtures for large scale imitation learning," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 145–164.

[296] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[297] H. Wang, C. Xiong, R. Wang, and X. Chen, "Bitvla: 1-bit vision-language-action models for robotics manipulation," *arXiv preprint arXiv:2506.07530*, 2025.

[298] K. Black, M. Y. Galliker, and S. Levine, "Real-time execution of action chunking flow policies," *arXiv preprint arXiv:2506.07339*, 2025.

[299] Y. Yue, Y. Wang, B. Kang, Y. Han, S. Wang, S. Song, J. Feng, and G. Huang, "Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution," *arXiv preprint arXiv:2411.02359*, 2024.

[300] S. Xu, Y. Wang, C. Xia, D. Zhu, T. Huang, and C. Xu, "Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation," *arXiv preprint arXiv:2502.02175*, 2025.

[301] T. Z. Zhao, V. Kumar, S. Levine, and C. 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] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 4066–4083. [Online]. Available: https://proceedings.mlr.press/v270/fu25b.html

[303] A. . Team, J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence *et al.*, "Aloha 2: An enhanced low-cost hardware for bimanual teleoperation," *arXiv preprint arXiv:2405.02292*, 2024.

[304] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 1910–1924.

[305] T. Buamanee, M. Kobayashi, Y. Uranishi, and H. Takemura, "Bi-act: Bilateral control-based imitation learning via action chunking with transformer," in *2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 2024, pp. 410–415.

[306] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. 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)*, 2024, pp. 12 156–12 163.

[307] Y. Qin, W. Yang, B. Huang, K. V. Wyk, H. Su, X. Wang, Y.-W. Chao, and D. 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] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[309] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. Van Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos *et al.*, "Curobo: Parallelized collision-free robot motion generation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8112–8119.

[310] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "Ace: A cross-platform and visual-exoskeletons system for low-cost dexterous teleoperation," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 4895–4911.

[311] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 2729–2749.

[312] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *arXiv preprint arXiv:2407.03162*, 2024.

[313] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. 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] T. L. Team, J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto, M. Z. Irshad *et al.*, "A careful examination of large behavior models for multitask dexterous manipulation," *arXiv preprint arXiv:2507.05331*, 2025.

[315] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. 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] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023.

[317] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiullah, "Robot utility models: General policies for zero-shot deployment in new environments," *arXiv preprint arXiv:2409.05865*, 2024.

[318] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and K. 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] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "Dexwild: Dexterous human interactions for in-the-wild robot policies," *arXiv preprint arXiv:2505.07813*, 2025.

[320] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote *et al.*, "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)*, June 2024, pp. 19 383–19 400.

[321] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol *et al.*, "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)*, June 2025, pp. 7061–7071.

[322] T. Perrett, A. Darkhalil, S. Sinha, O. Emara, S. Pollard, K. Parida, K. Liu, P. Gatti, S. Bansal, K. Flanagan *et al.*, "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] Z. Lv, N. Charron, P. Moulon, A. Gamino, C. Peng, C. Sweeney, E. Miller, H. Tang, J. Meissner, J. Dong *et al.*, "Aria everyday activities dataset," *arXiv preprint arXiv:2402.13349*, 2024.

[324] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," *arXiv preprint arXiv:2410.24221*, 2024.

[325] V. Liu, A. Adeniji, H. Zhan, S. Haldar, R. Bhirangi, P. Abbeel, and L. Pinto, "Egozero: Robot learning from smart glasses," *arXiv preprint arXiv:2505.20290*, 2025.

[326] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen *et al.*, "Humanoid policy human policy," *arXiv preprint arXiv:2503.13441*, 2025.

[327] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, 29–31 Oct 2018, pp. 879–893.

[328] K. Crowston, "Amazon mechanical turk: A research tool for organizations and information systems scholars," in *Shaping the Future of ICT Research. Methods and Approaches*, A. Bhattacherjee and B. Fitzgerald, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 210–221.

[329] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," *IEEE Robotics and Automation Letters (RA-L)*, pp. 1–8, 2023.

[330] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[331] Q. Sun, P. Hong, T. D. Pala, V. Toh, U.-X. Tan, D. Ghosal, and S. Poria, "Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 14 199–14 214. [Online]. Available: https://aclanthology.org/2025.acl-long.695/

[332] N. Blank, M. Reuss, M. Rühle, Ö. E. Yağmurlu, F. Wenzel, O. Mees, and R. Lioutikov, "Scaling robot policy learning via zero-shot labeling with foundation models," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 4158–4187. [Online]. Available: https://proceedings.mlr.press/v270/blank25a.html

[333] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang *et al.*, "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," *arXiv preprint arXiv:2412.13877*, 2025.

[334] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[335] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. 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] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, 29–31 Oct 2018, pp. 651–673.

[337] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," *arXiv preprint arXiv:2104.08212*, 2021.

[338] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," in *Proceedings of the Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 30 Oct–01 Nov 2020, pp. 885–897.

[339] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. 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] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 1723–1736. [Online]. Available: https://proceedings.mlr.press/v229/walke23a.html

[341] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 991–1002.

[342] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 3, pp. 7327–7334, 2022.

[343] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 44 776–44 791.

[344] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. 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)*, June 2022, pp. 21 013–21 022.

[345] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. 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)*, June 2024, pp. 445–456.

[346] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2o: Two hands manipulating objects for first person interaction recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 138–10 148.

[347] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "Arctic: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[348] Y. Li, Z. Cao, A. Liang, B. Liang, L. Chen, H. Zhao, and C. 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] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[350] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. 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*, 2017, pp. 5842–5850.

[351] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2022.

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

[353] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 1820–1864. [Online]. Available: https://proceedings.mlr.press/v229/mandlekar23a.html

[354] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," *arXiv preprint arXiv:2410.24185*, 2025.

[355] P. Sharma, L. Mohan, L. Pinto, and A. Gupta, "Multiple interactions made easy (mime): Large scale demonstrations data for imitation," in *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, 29–31 Oct 2018, pp. 906–915.

[356] S. Ramos, S. Girgin, L. Hussenot, D. Vincent, H. Yakubovich, D. Toyama, A. Gergely, P. Stanczyk, R. Marinier, J. Harmsen *et al.*, "Rlds: an ecosystem to generate, share and use datasets in reinforcement learning," *arXiv preprint arXiv:2111.02767*, 2021.

[357] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard, "Latent plans for task-agnostic offline reinforcement learning," in *Proceedings of The 6th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 1838–1849. [Online]. Available: https://proceedings.mlr.press/v205/rosete-beas23a.html

[358] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 576–11 582.

[359] S. Dass, J. Yapeter, J. Zhang, J. Zhang, K. Pertsch, S. Nikolaidis, and J. J. Lim, "Clvr jaco play dataset," 2023. [Online]. Available: https://github.com/clvrai/clvr_jaco_play_dataset

[360] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine, "Multistage cable routing through hierarchical imitation learning," *IEEE Transactions on Robotics*, vol. 40, pp. 1476–1491, 2024.

[361] L. Y. Chen, S. Adebola, and K. Goldberg, "Berkeley UR5 demonstration dataset," https://sites.google.com/view/berkeley-ur5/home.

[362] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto *et al.*, "Train offline, test online: A real robot learning benchmark," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9197–9203.

[363] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 4788–4795.

[364] N. Hirose, D. Shah, A. Sridhar, and S. Levine, "Sacson: Scalable autonomous control for social navigation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 1, pp. 49–56, 2024.

[365] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. 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. 11 807–11 814, 2022.

[366] D. Shah, B. Eysenbach, N. Rhinehart, and S. Levine, "Rapid exploration for open-world navigation with latent goal models," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 674–684. [Online]. Available: https://proceedings.mlr.press/v164/shah22a.html

[367] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[368] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 645–652.

[369] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran, and V. Kumar, "Cacti: A framework for scalable multi-task multi-scene visual imitation learning," in *CoRL 2022 Workshop on Pre-training Robot Learning*, 2022.

[370] Q. Chen, S. C. Kiami, A. Gupta, and V. 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] T. Yu, T. Xiao, J. Tompson, A. Stone, S. Wang, A. Brohan, J. Singh, C. Tan, D. M, J. Peralta *et al.*, "Scaling robot learning with semantically imagined experience," in *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 2023.

[372] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut *et al.*, "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)*, June 2023, pp. 18 359–18 369.

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

[374] T. Xiao, H. Chan, P. Sermanet, A. Wahid, A. Brohan, K. Hausman, S. Levine, and J. 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] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 627–635.

[376] L. Ke, Y. Zhang, A. Deshpande, S. Srinivasa, and A. Gupta, "Ccil: Continuity-based data augmentation for corrective imitation learning," in *International Conference on Learning Representations (ICLR)*, 2024.

[377] M. Iskandar, C. Ott, O. Eiberger, M. Keppler, A. Albu-Schäffer, and A. Dietrich, "Joint-level control of the dlr lightweight robot sara," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8903–8910.

[378] S. Guist, J. Schneider, H. Ma, L. Chen, V. Berenz, J. Martus, H. Ott, F. Grüninger, M. Muehlebach, J. Fiene *et al.*, "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] K. Shaw, A. Agarwal, and D. 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] M. H. Khan, S. Asfaw, D. Iarchuk, M. A. Cabrera, L. Moreno, I. Tokmurziyev, and D. Tsetserukou, "Shake-vla: Vision-language-action model-based system for bimanual robotic manipulations and liquid mixing," *arXiv preprint arXiv:2501.06919*, 2025.

[381] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, Y. Zhu, and K. Lin, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv preprint arXiv:2009.12293*, 2020.

[382] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 1678–1690. [Online]. Available: https://proceedings.mlr.press/v164/mandlekar22a.html

[383] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "RoboCasa: Large-Scale Simulation of Household Tasks for Generalist Robots," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[384] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Proceedings of the Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, L. P.

Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 30 Oct–01 Nov 2020, pp. 1094–1100.

[385] H. Xue, X. Huang, D. Niu, Q. Liao, T. Kragerud, J. T. Gravdahl, X. B. Peng, G. Shi, T. Darrell, K. Screenath *et al.*, "Leverb: Humanoid whole-body control with latent vision-language instruction," *arXiv preprint arXiv:2506.13751*, 2025.

[386] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021.

[387] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *International Conference on Learning Representations (ICLR)*, 2023.

[388] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan *et al.*, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv preprint arXiv:2410.00425*, 2024.

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

[390] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu *et al.*, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 27 649–27 660.

[391] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Q. Liang, Z. Li, X. Lin, Y. Ge, Z. 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] S. Zhang, P. Wicke, L. K. Şenel, L. Figueredo, A. Naceri, S. Haddadin, B. Plank, and H. Schütze, "Lohoravens: A long-horizon language-conditioned benchmark for robotic tabletop manipulation," *arXiv preprint arXiv:2310.12020*, 2023.

[393] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[394] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 251–266.

[395] X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min *et al.*, "Habitat 3.0: A co-habitat for humans, avatars and robots," *arXiv preprint arXiv:2310.13724*, 2023.

[396] S. James, Z. Ma, D. R. Arrojo, and A. 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] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: A benchmark for evaluating generalization for robotic manipulation," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024.

[398] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv*, 2017.

[399] K. Ehsani, T. Gupta, R. Hendrix, J. Salvador, L. Weihs, K.-H. Zeng, K. P. Singh, Y. Kim, W. Han, A. Herrasti *et al.*, "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)*, June 2024, pp. 16 238–16 250.

[400] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani *et al.*, "Evaluating real-world robot manipulation policies in simulation," in *Proceedings of The 8th Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 3705–3728.

[401] P. Atreya, K. Pertsch, T. Lee, M. J. Kim, A. Jain, A. Kuramshin, C. Eppner, C. Neary, E. Hu, F. Ramos *et al.*, "Roboarena: Distributed real-world evaluation of generalist robot policies," *arXiv preprint arXiv:2506.18123*, 2025.

[402] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *arXiv preprint arXiv:2306.03310*, 2023.

[403] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar *et al.*, "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] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[405] L. Zheng, F. Yan, F. Liu, C. Feng, Z. Kang, and L. Ma, "Robocas: A benchmark for robotic manipulation in complex object arrangement scenarios," in *NeurIPS: Datasets and Benchmarks Track*, 2024.

[406] C. Bao, H. Xu, Y. Qin, and X. Wang, "Dexart: Benchmarking generalizable dexterous manipulation with articulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 190–21 200.

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

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

[409] S. James, M. Freese, and A. J. Davison, "Pyrep: Bringing v-rep to deep robot learning," *arXiv preprint arXiv:1906.11176*, 2019.

[410] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford *et al.*, "Robothor: An open simulation-to-real embodied ai platform," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[411] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 5982–5994.

[412] Z. Wang, Z. Zhou, J. Song, Y. Huang, Z. Shu, and L. Ma, "Vlatest: Testing and evaluating vision-language-action models for robotic manipulation," *arXiv preprint arXiv:2409.12894*, 2024.

[413] T. Wang, C. Han, J. C. Liang, W. Yang, D. Liu, L. X. Zhang, Q. Wang, J. Luo, and R. Tang, "Exploring the adversarial vulnerabilities of vision-language-action models in robotics," *arXiv preprint arXiv:2411.13587*, 2024.

[414] H. Cheng, E. Xiao, C. Yu, Z. Yao, J. Cao, Q. Zhang, J. Wang, M. Sun, K. Xu, J. Gu *et al.*, "Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models," *arXiv preprint arXiv:2409.13174*, 2024.

[415] H. Lu, H. Li, P. S. Shahani, S. Herbers, and M. Scheutz, "Probing a vision-language-action model for symbolic states and integration into a cognitive architecture," *arXiv preprint arXiv:2502.04558*, 2025.

[416] H.-T. L. Chiang, Z. Xu, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah *et al.*, "Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs," *arXiv preprint arXiv:2407.07775*, 2024.

[417] V. Serpiva, A. Lykov, A. Myshlyaev, M. H. Khan, A. A. Abdulkarim, O. Sautenkov, and D. Tsetserukou, "Racevla: Vla-based racing drone navigation with human-like behaviour," *arXiv preprint arXiv:2503.02572*, 2025.

[418] A. Lykov, V. Serpiva, M. H. Khan, O. Sautenkov, A. Myshlyaev, G. Tadevosyan, Y. Yaqoot, and D. 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] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," *arXiv preprint arXiv:2408.11812*, 2024.

[420] R. Zheng, Y. Liang, S. Huang, J. Gao, H. D. III, A. Kolobov, F. Huang, and J. Yang, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," *arXiv preprint arXiv:2412.10345*, 2024.

[421] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu *et al.*, "Egovla: Learning vision-

language-action models from egocentric human videos," *arXiv preprint arXiv:2507.12440*, 2025.
[422] F. Nolte, B. Schölkopf, and I. Posner, "Is single-view mesh reconstruction ready for robotics?" *arXiv preprint arXiv:2505.17966*, 2025.
[423] Q. Gu, Y. Ju, S. Sun, I. Gilitschenski, H. Nishimura, M. Itkina, and F. Shkurti, "Safe: Multitask failure detection for vision-language-action models," *arXiv preprint arXiv:2506.09937*, 2025.
[424] Z. Yang, Y. Chen, X. Zhou, J. Yan, D. Song, Y. Liu, Y. Li, Y. Zhang, P. Zhou, H. Chen *et al.*, "Agentic robot: A brain-inspired framework for vision-language-action models in embodied agents," *arXiv preprint arXiv:2505.23450*, 2025.
[425] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse control tasks through world models," *Nature*, pp. 1–7, 2025.