

Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey

Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, Liqiang Nie

Abstract—Robotic manipulation, as a critical frontier in robotics and embodied AI, demands precise motor control and integrated understanding of visual and semantic cues in dynamic environments. Traditional approaches, grounded in predefined task specifications and rigid control policies, often struggle to scale or generalize in unstructured, novel scenarios. In recent years, Vision-Language-Action (VLA) models, built upon Large Vision-Language Models (VLMs) pretrained on vast image-text datasets, have emerged as a transformative paradigm. By leveraging large VLMs' capabilities in open-world generalization, hierarchical task planning, knowledge-augmented reasoning, and rich multimodal fusion, these models empower robots to interpret high-level instructions, recognize unseen environments and execute complex manipulation tasks. This survey provides the first systematic, taxonomy-oriented review of large VLM-based VLA models for robotic manipulation. We begin by clearly defining large VLM-based VLA models and delineating two principal architectural paradigms: (1) **monolithic models**, encompassing single-system and dual-system designs with differing levels of integration; and (2) **hierarchical models**, which explicitly decouple planning from execution via interpretable intermediate representations. Building on this foundation, we present an in-depth examination of large VLM-based VLA models: (1) integration with advanced domains, including reinforcement learning, training-free optimization, learning from human videos, and world model integration; (2) synthesis of distinctive characteristics, consolidating architectural traits, operational strengths, and the datasets and benchmarks that support their development; (3) identification of promising directions, including memory mechanisms, 4D perception, efficient adaptation, multi-agent cooperation, and other emerging capabilities. This survey consolidates recent advances to resolve inconsistencies in existing taxonomies, mitigate research fragmentation, and fill a critical gap through the systematic integration of studies at the intersection of large VLMs and robotic manipulation. We provide a regularly updated project page to document ongoing progress: <https://github.com/JiuTian-VL/Large VLM-based VLA for Robotic Manipulation>.

Index Terms—Vision-Language-Action Models, Robotic Manipulation, Embodied AI, Large Vision-Language Models

1 INTRODUCTION

ROBOTIC manipulation stands as a pivotal challenge at the confluence of robotics and embodied artificial intelligence [1]–[5]. Its realization necessitates not only precise motor control but also a profound comprehension of diverse visual and semantic cues within complex, dynamic environments. The widespread utility of robotic manipulation is evident across numerous domains, including advanced manufacturing, efficient logistics, precision healthcare, and versatile domestic service [6]–[8]. Traditional manipulation methods [9]–[16] have been predominantly anchored in meticulously engineered control policies and rigidly predefined task specifications. However, these approaches demonstrably falter in unstructured, real-world settings—particularly when confronted with novel objects, ambiguous natural language instructions, or previously unseen environmental configurations—thereby exposing inherent limitations in their scalability and generalization capabilities.

Recently, large Vision-Language Models (VLMs) [17]–[25] have emerged as a transformative paradigm. Pretrained on vast web-scale image-text datasets, large VLMs exhibit a remarkable aptitude for bridging the semantic chasm between visual perception and natural language understanding. This innovative capacity enables large VLMs to

Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen and Liqiang Nie are with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China, 518055 (e-mail: shaorui@hit.edu.cn, {liwei2024, zls030726, zhangrenshan, 25s151095, 220310324}@stu.hit.edu.cn, nieliqiang@gmail.com).

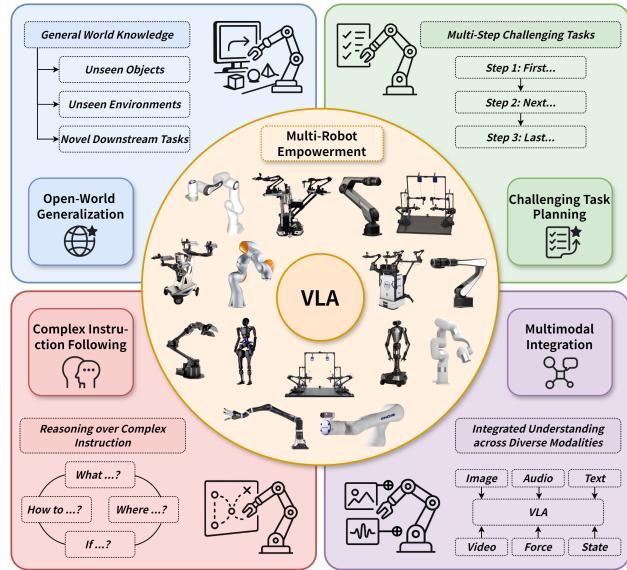


Fig. 1: Illustration of core advantages of large VLM-based Vision-Language-Action (VLA) models for robotic manipulation. Large VLM-based VLA models leverages the strengths of large Vision-Language Models (VLMs), including (1) open-world generalization, (2) hierarchical task planning, (3) knowledge-augmented reasoning, and (4) rich multimodal fusion. These capabilities empower diverse robotic arms and significantly enhance robotic intelligence.

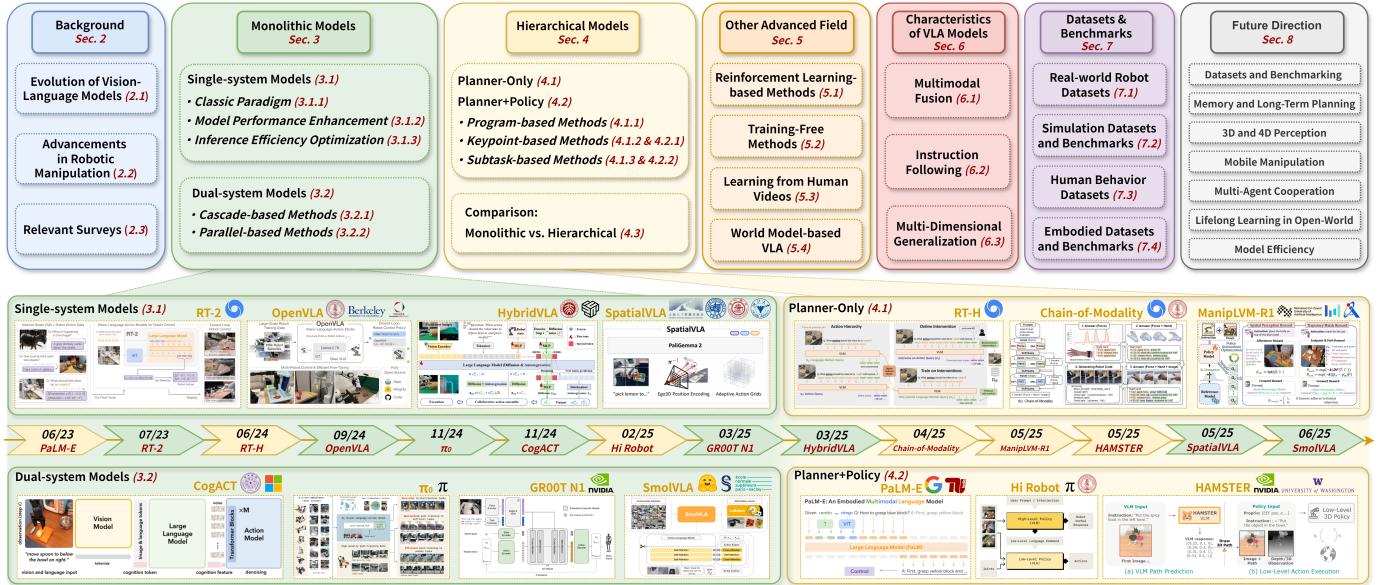


Fig. 2: Outline of the organization of our comprehensive survey (top) and a chronological timeline of notable developments in large VLM-based Vision-Language-Action models for robotic manipulation (bottom). The timeline highlights representative milestones for both monolithic models and hierarchical models, providing a perspective on the field’s recent progress.

interpret complex visual scenes in conjunction with textual descriptions, moving beyond mere object recognition to encompass a holistic contextual understanding. The subsequent integration of large VLMs into robotic systems has catalyzed the development of a novel class of models: large VLM-based Vision-Language-Action (VLA) models [26]–[32]. As illustrated in Fig. 1, this emerging paradigm offers great potential to overcome the fundamental limitations of conventional robotic pipelines. It enables robots to interpret high-level human instructions, generalize to unseen objects and scenarios, reason about complex spatial relationships, and perform sophisticated manipulation tasks in dynamic, unstructured environments. For instance, a VLA model could execute commands such as “place the red mug next to the laptop onto the top shelf,” a task demanding a sophisticated fusion of visual grounding, spatial reasoning, and sequential motion planning.

In this work, based on an extensive review of recent research [26]–[37] and a refined understanding of the field [38]–[43], we propose a coherent definition of large VLM-based VLA models: the model (1) leverages a large VLM to understand visual observations and natural language instructions, and (2) performs reasoning processes that directly or indirectly serve robotic action generation. We further distinguish two principal categories of large VLM-based VLA models, as shown in Fig. 2 and Fig. 3:

- **Monolithic Models** (Fig. 3, left) comprise single-system and dual-system implementations. (1) Single-system models [26], [27], [44], [45] integrate both environmental comprehension (including visual perception, linguistic understanding, and robot state awareness) and action generation within a unified architecture. In contrast, (2) dual-system models [29]–[32] employ a VLM backbone for scene interpretation and an action expert for action determination, exchanging information via the propagation of latent representations.

• **Hierarchical Models** (Fig. 3, right) [46]–[50] explicitly decouple planning from policy execution. These differ from dual-system end-to-end approaches through two defining characteristics: (1) Structured intermediate outputs where planner modules generate interpretable representations such as keypoint detections, affordance maps, or trajectory proposals, which are then processed by policy modules to formulate executable actions; and (2) Decoupled training paradigms that enable independent optimization of hierarchical modules through specialized loss functions or API-mediated interactions.

This taxonomy emphasizes critical design dimensions in VLA development, particularly regarding system integration granularity and the explicitness of cognitive decomposition, while maintaining essential connections to modern representation learning paradigms.

Operating under the above definitions and taxonomy, our comprehensive survey across the broad spectrum of related work reveals several critical gaps within the nascent VLA field, with the organization of this survey illustrated in Fig. 2. First, the terminology and modeling assumptions in this field remain inconsistent, and the research landscape is fragmented across disciplines (robotics, computer vision, natural language processing, etc.). Second, existing reviews tend to focus either on VLMs [51]–[55] or on robotic manipulation in isolation [2], [56]–[59], lacking a comprehensive synthesis of the unique challenges and advances that arise at their intersection. Therefore, there is a pressing need for a systematic and principled survey that elucidates the foundations of large VLM-based VLA models, organizes the space of relevant methods, and outlines future directions for this integrated paradigm. This survey aims to address this gap. We present a structured and in-depth overview of advances in large VLM-based VLA models research, aiming to provide a panoramic perspective of the field to foster a deeper understanding and drive future breakthroughs. Our

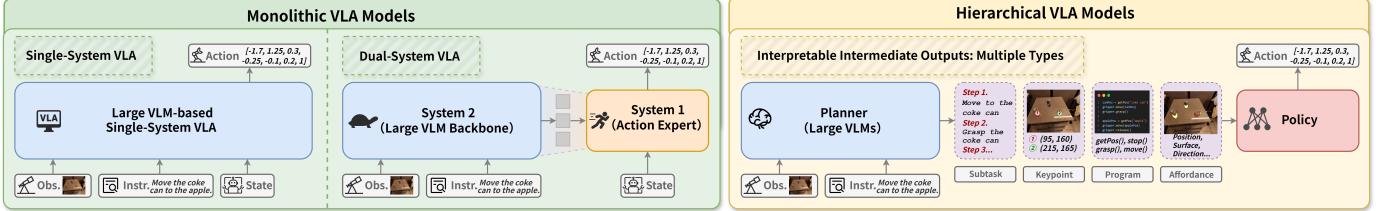


Fig. 3: Comparison of the two principal categories of large VLM-based VLA models. Monolithic models (Sec. 3) integrate perception, language understanding, and action generation within single- or dual-system architectures, with the latter incorporating an additional action expert. In contrast, hierarchical models (Sec. 4) decouple planning from policy execution through interpretable intermediate outputs (e.g., subtasks, keypoints, programs, affordances).

key contributions are summarized as follows:

- **A Longitudinal Synthesis of large VLM-based VLA Models Development:** We systematically review the evolutionary trajectory of VLMs, the technical advancements in manipulation learning, and the subsequent emergence of the large VLM-based VLA paradigm. Furthermore, we examine the development of monolithic models and hierarchical models, identifying key challenges and outlining future directions.
- **A Cross-Cutting Synthesis of large VLM-based VLA Modeling Practices:** We provide a more fine-grained comparative taxonomy of monolithic models and hierarchical models, examining them in detail from both structural and functional perspectives. We further explore advanced research frontiers of large VLM-based VLA models, highlighting their distinctive characteristics and the datasets that underpin their development.

This synthesis provides a high-level summary and a conceptual roadmap for understanding the field's development and structural organization.

The remainder of this survey is organized as follows. As illustrated in Fig. 2, Sec. 2 presents essential background knowledge on the evolution of VLMs and the foundational aspects of robotic manipulation. Sec. 3 examines monolithic models, detailing single-system and dual-system architectures, their respective advantages, and design trade-offs. Sec. 4 explores hierarchical models, categorizing them into planner-only and planner-policy frameworks, and further sub-classifying by intermediate representation types such as subtasks, keypoints, and programs. Sec. 5 discusses other advanced fields, including RL-based optimization, training-free methods, learning from human videos, and world model-based approaches. Sec. 6 analyzes the defining characteristics of large VLM-based VLA models, covering multimodal fusion, instruction following, and multi-dimensional generalization. Sec. 7 categorizes and analyzes the diverse datasets and benchmarks employed in large VLM-based VLA models research, spanning simulated, real-world, and human interaction data. Sec. 8 explores critical open challenges and promising future research directions. Finally, Sec. 9 concludes the survey.

2 BACKGROUND

2.1 Evolution of Vision-Language Models

Recently, the emergence of large VLMs [17]–[25] has been marked by a shift from task-specific architectures to uni-

fied frameworks capable of handling diverse multimodal tasks [60]–[64]. Modern VLMs, such as LLaVA1.5 [18] and Qwen-VL [20], typically adopt a three-component architecture: a vision encoder to encode visual input, a projector to align visual features with textual embeddings, and a large language model to mediate multimodal reasoning. This design unifies multimodal tasks as textual outputs. It enables VLMs to perform not only traditional tasks like visual question answering or object detection, but also more advanced capabilities, including compositional reasoning and spatial navigation, which are critical for real-world applications. Building on this foundation, DriveVLM [65] demonstrates the integration of VLMs into autonomous driving systems, while CogAgent [66] shows VLMs' ability to interact with graphical user interfaces. These advancements highlight the growing sophistication of large VLMs in bridging vision and language, while also revealing the emerging potential of using large VLMs to build broad real-world applications.

Crucially, the potential of large VLMs in real-world applications [67]–[76] rests on their generalization capabilities afforded by visual instruction tuning. By training on large-scale, carefully curated visual instruction-following datasets, VLMs acquire the flexibility to comprehend abstract or open-ended instructions and to generalize to unseen multimodal scenarios. LLaVA [21] pioneers this paradigm by leveraging GPT-4 to convert raw image-text pairs into 150K conversational samples, significantly enhancing its ability to address open-ended queries. InternV2 [22] advances by introducing an LLM-guided automated filtering pipeline. It helps clean noisy annotations, reducing anomalous behaviors of the model. Collectively, these advances demonstrate how visual instruction tuning prepares large VLMs to follow high-level commands. This capability forms the foundation for extending large VLMs to VLA for robotic manipulation, which requires robust reasoning across vision, language, and action in the real world with ambiguous instructions.

Recent advancements in VLMs further emphasize scalability for enhanced multimodal perception and reasoning. Models like LLaVA-OneVision [77] unify image and video processing within a single framework. Qwen-2VL [78] introduces dynamic resolution support for arbitrary input sizes, while Vision-R1 [79] leverages reinforcement learning to improve chain-of-thought reasoning. Collectively, these innovations empower VLMs with stronger real-world perception and decision-making capabilities, laying the groundwork for VLMs to further integrate action generation.

Building on these breakthroughs, the next frontier for large VLMs is to transition from passive observation to active interaction with physical environments. While models like DriveVLM [65] and CogAgent [66] already demonstrate VLMs' ability to process real-world inputs and generate high-level plans, they remain unable to engage directly with the real world in an embodied, physical manner. Equipping VLMs with robotic manipulation capabilities thus emerges as a particularly promising avenue for future research.

2.2 Advancements in Robotic Manipulation

Early language-conditioned robotic systems typically pair separate vision encoders with language modules or planners, rather than using a unified multimodal model. For example, CLIPort [80] employs the CLIP vision-language encoder [81] for semantic grounding and a Transporter [82] network for spatial reasoning. Likewise, RT-1 [83] uses a CNN-based visual encoder [84] conditioned by a separate language embedding derived from a universal sentence encoder to map camera images and text instructions to motor actions. By integrating separate pretrained visual and language encoders with multi-task robotic manipulation data, these early methods achieve a certain degree of multi-task generalization. However, their modular designs struggle to generalize to unseen concepts and face difficulties when handling complex or ambiguous instructions.

In contrast, the latest VLA models tightly integrate vision, language, and action control in a unified system. The RT-2 [27] model exemplifies this shift. RT-2 [27] begins with a pretrained large VLM backbone (e.g., PaLM-E [85] or PaLI-X [86]) and co-trains it on both internet-scale vision-language tasks and real robot trajectories. Crucially, robot actions are cast as text tokens and included in the same training corpus as normal language outputs. This simple scheme allows the model to “absorb” robot control as another language task, resulting in a unified VLA model. Empirically, RT-2 shows a dramatically stronger semantic understanding ability. Compared to RT-1 [83], RT-2 [27] generalizes to novel objects and unseen instructions (e.g., placing an object on a particular number) and can perform basic reasoning (choosing the smallest or closest object). In other words, the large-scale vision-language pretraining imbues the robotic policy with real-world knowledge and stronger language grounding that earlier models lack.

Following RT-2's demonstration of how Internet-scale VLM pretraining can boost robotic control, more VLA models have emerged. For example, π_0 [29] adopts a flow-matching architecture on top of a pretrained VLM, trained on diverse dexterous robot datasets, which yields strong zero-shot generalization and easy adaptation to new tasks via fine-tuning. In addition, OpenVLA [26] has been released as a 7B open-source VLA model pretrained on approximately 970k real-world robot demonstrations. It achieves superior generalist manipulation performance and supports efficient fine-tuning on consumer hardware via techniques like LoRA [87]. Together, these advancements underscore a trend toward greater generality and open accessibility in VLA-driven robotic manipulation.

2.3 Relevant Surveys

The growing interest in VLA models for embodied AI has inspired several surveys [1]–[6], [8], [88], [89], yet most focus on broader architectural paradigms or diverse application domains, leaving a gap in the systematic exploration of large VLM-based VLA systems specifically tailored for robotic manipulation. For example, Ma et al. [2] conduct a comprehensive survey of VLA architectures, reviewing modular, end-to-end, and hybrid approaches for integrating vision, language, and action modalities. However, this survey lacks a focused analysis of recent trends that leverage pre-trained VLMs as foundational components. These models have emerged as a dominant paradigm due to their strong cross-modal alignment and zero-shot generalization capabilities. Sapkota et al. [8] provide an extensive investigation into a broader range of VLA applications, such as autonomous driving, augmented reality navigation, and precision agriculture. However, its breadth dilutes the depth required for robotics-specific challenges such as real-time actuation constraints, sensor noise robustness, and long-term decision-making. A survey by Wang et al. [88] offers an early overview of the integration of text-only LLMs into robotic task planning. It emphasizes their capability to generate precise action plans from natural language instructions. However, the study primarily focuses on high-level reasoning and does not address the challenge of grounding VLMs in robotic visual perception and action determination, which is a gap that recent VLA models are designed to fill.

To address the lack of a comprehensive and in-depth survey in this emerging field, we provide a structured overview of recent VLA research. Building on the definition of VLA models described in Sec. 1, we trace the development of VLAs in robotics, analyze key architectures and learning paradigms, and review relevant datasets and benchmarks. Our aim is to offer a comprehensive perspective that supports deeper understanding and drives future advances.

3 MONOLITHIC MODELS

Monolithic VLA models are mainly implemented in two ways: single-system and dual-system architectures, as illustrated in Fig. 3 left. In the single-system design (Sec. 3.1), visual perception, language instructions, and robot states are jointly fed into a unified model that processes all modalities and decodes executable actions via autoregressive or parallel decoding. In contrast, dual-system architectures (Sec. 3.2) divide functionality into two cooperating modules: System 2 (VLM backbone) performs slower but more generalized reflective reasoning, while System 1 (action expert) focuses on fast processing to support reactive behaviors. Single-system VLAs offer architectural simplicity, streamlined development, and avoidance of complex inter-module communication while dual-system VLAs leverage a division of labor to combine reactive speed with deliberate accuracy.

3.1 Single-system Models

Single-system VLA models embody the monolithic design philosophy. They aim to transfer the semantic knowledge of large VLMs to robotic manipulation tasks through a unified model. This section reviews the paradigms of single-system

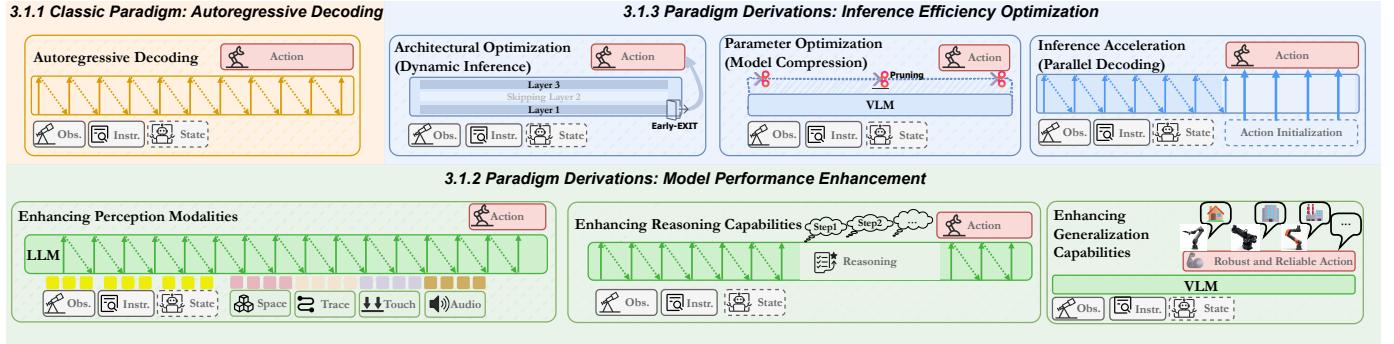


Fig. 4: Comparison of representative paradigms in monolithic single-system models. Sec. 3.1.1 illustrates the model schematic of classical autoregressive decoding. Arrows denote the flow of information. Sec. 3.1.2 presents approaches to enhance model capabilities by incorporating additional modalities, leveraging chain-of-thought reasoning, and strengthening generalization. Sec. 3.1.3 discusses some methods for improving the inference efficiency through architectural refinement, parameter design, and decoding strategies.

models and explores two main research directions: enhancing the model’s capabilities to solve complex tasks and improving inference efficiency for practical deployment. All models are summarized in Tab. 1.

3.1.1 Classic Paradigm: Autoregressive Decoding

The autoregressive decoding paradigm draws directly from the sequence generation capability of LLMs. By discretizing the robot’s continuous action space into a token sequence, the model can sequentially predict action tokens. As shown in “Autoregressive Decoding” of Fig. 4, the VLM receives the visual observation, natural language instruction, and optionally the robot state as input. Then it autoregressively generates action tokens, which can be converted into executable actions through a downstream de-tokenizer.

The RT series [27], [83], [90] and OpenVLA [26] are typical examples of this paradigm. RT-1 [83] introduces Transformer [91] into the VLA domain and encodes continuous robot actions as discrete tokens. RT-2 [27] inherits this token-based action representation and applies it to a larger-scale VLM. It co-fine-tunes on internet-scale vision-language and robot trajectory data, successfully transferring rich visual-semantic knowledge from web data to robotic manipulation tasks. RT-2-X [90] further improves cross-robot skill transfer by applying co-fine-tuning on the Open X-Embodiment (OXE) dataset [90]. Going a step further, OpenVLA [26] replaces the large-parameter vision encoder in RT series with a combination of SigLIP [92] and DINOv2 [93]. Through fine-tuning on large-scale real-world robotic manipulation data, it achieves superior performance with fewer model parameters. Its fully open-sourced nature positions it as a widely used baseline in subsequent research.

3.1.2 Model Performance Enhancement

The development of VLA models has been accompanied by improvements in model performance. To better align model capabilities with human intelligence, performance enhancement efforts have focused on three directions:

Enhancing Perception Modalities. As shown in Fig. 4, these models improve real-world information acquisition by expanding input modalities. **(1) 3D perception.** Since the

real world is three-dimensional and continuous, large VLM-based VLA models cannot fully understand it by relying solely on 2D images. Accordingly, Leo Agent [94] directly obtains object-centric 3D point clouds from simulation environments and encodes them using PointNet++ [119]; the encoded features are then fed into a spatial transformer [120] to generate “Space” information. In real-world settings, SpatialVLA [100] proposes Ego3D position encoding, which estimates depth from 2D images and back-projects it into egocentric 3D coordinates, fusing pixel-wise 3D positions with 2D features to produce egocentric “Space” information. Similarly, BridgeVLA [35] reconstructs 3D point clouds from RGB-D images and generates three orthographic projection views, enabling “Space” information to be represented in a 2D format compatible with large VLMs.

(2) 4D perception. To better capture both spatial structure and temporal dynamics in robotic manipulation, 4D perception is required for more comprehensive scene understanding. TraceVLA [97] overlays sampled motion-point trajectories on the current image to create the “Trace” information in Fig. 4. Combined with the original observation, it equips the large VLM with spatiotemporal understanding. Furthermore, 4D-VLA [108] integrates 3D coordinates into visual features to resolve spatial coordinate inconsistency, and leverages memory bank sampling of historical keyframes to mitigate temporal state ambiguity, thereby improving spatiotemporal reasoning. ST-VLA [110] introduces the “Spatial Traces” approach, which combines historical temporal “Trace” information with spatial depth information of the scene, surpassing prior SpatialVLA [100] and TraceVLA [97] models in certain robotic manipulation tasks.

(3) Tactile and auditory perception. VTLA [103] encodes tactile information via a vision encoder to align it with visual sequences, and then feeds it together with visual and textual tokens into the LLM, thereby constructing a “Vision-Tactile-Language-Action model”. VLAs [102] employs the “Whisper Encoder” to extract speech features and projects them into the LLM embedding space through an MLP, thereby introducing the “Audio” information in Fig. 4. Both VTLA and VLAs construct task-specific datasets for fine-tuning. Moreover, FuSe [98] learns to align “Touch” and “Audio” modalities with language concepts during fine-

TABLE 1: Single-system VLA models. In the LLM / VLM column, omission of the V-Encoder indicates a VLM; otherwise, it represents an LLM. In the Learning column, “AD” denotes Autoregressive Decoding and “PD” denotes Parallel Decoding. “SFT” denotes fine-tuning distinct from action-prediction imitation learning, where tasks like captioning, VQA, reasoning and others all qualify as SFT. “A” and “B” in parentheses represent the learning methods used by Action head or Backbone.

Model	V-Encoder	LLM / VLM	Learning	Contribution
Classic Paradigm: Autoregressive Decoding				
RT-2 [27]	-	PaLI-X / PaLM-E	AD (A), SFT (B)	Represent actions as VLM tokens to enable generalization.
RT-2-X [90]	ViT-22B	UL2	AD (A), SFT (B)	Fine-tune on cross-robot data for positive skill transfer.
OpenVLA [26]	DINOv2 + SigLIP	LLaMA2-7B	AD (A)	Open-source 7B-parameter VLA model for generalist robot control.
Paradigm Derivations: Model Performance Enhancement				
LEO agent [94]	ConvNext	Vicuna-7B	AD (A), SFT (B)	Combine object-centric 3D features with LLM for action.
ECoT [95]	DINOv2 + SigLIP	LLaMA2-7B	AD (A), SFT (B)	Incorporate chain-of-thought to enhance policy explainability.
ReVLA [96]	DINOv2 + SigLIP	LLaMA2-7B	AD (A)	Reverse backbone gradually to preserve visual generalization.
TraceVLA [97]	DINOv2 + SigLIP	LLaMA2-7B	AD (A)	Propose visual trace prompting for spatiotemporal awareness.
FuSe [98]	-	PaliGemma-3B	AD (A), SFT (B)	Leverage natural language for cross-modal fine-tuning.
UniAct [99]	-	LLaVA-0.5B	PD (A)	Propose universal action space for versatile and adaptive control.
SpatialVLA [100]	SigLIP	Gemma2	AD (A)	Improve generalization via 3D encoding and action grid.
UP-VLA [101]	ViT + VQ-GAN	Phi1.5-1.3B	AD (A), SFT (B)	Propose unified training for semantic-spatial understanding.
VLAS [102]	CLIP	Vicuna-7B	AD (A), SFT (B)	Introduce voice modality to VLA and construct a paired dataset.
HybridVLA [34]	DINOv2 + SigLIP	LLaMA2-7B	Diff., AD (A)	Integrate diffusion and autoregressive policies to improve success.
CoT-VLA [33]	-	VILA-U	AD+PD(A), SFT(B)	Propose a visual chain-of-thought to improve planning.
VTLA [103]	-	Qwen2-VL-7B	AD (A)	Integrate visual and tactile inputs to improve task success.
OE-VLA [104]	SigLIP	Qwen1.5-7B	AD (A), SFT (B)	Introduce four open-ended tasks to expand interaction modalities.
ReFineVLA [105]	SigLIP	Gemma2	AD (A), SFT (B)	Propose reasoning-aware framework to fine-tune VLAs effectively.
LoHoVLA [106]	SigLIP	Gemma-2B	AD (A), SFT (B)	Address long-horizon tasks via hierarchical closed-loop control.
BridgeVLA [35]	SigLIP	Gemma	PD (A), SFT (B)	Project 3D data into 2D space for efficient action prediction
UnifiedVLA [107]	-	Emu3	AD (A), SFT (B)	Convert all input signals into tokens to build a unified model.
WorldVLA [38]	-	Chameleon	AD (A), SFT (B)	Combine world and action models for bidirectional improvement.
4D-VLA [108]	-	InternVL-4B	PD (A)	Integrate 4D spatiotemporal cues for efficient VLA pretraining.
VOTE [109]	DINOv2 + SigLIP	LLaMA2-7B	PD (A)	Introduce voting strategy to increase action prediction accuracy.
ST-VLA [110]	-	PaliGemma2	AD (A), SFT (B)	Project visual traces onto depth maps for better understanding.
Paradigm Derivations: Inference Efficiency Optimization				
RoboFlamingo [111]	ViT	MPT-1B	PD (A), SFT (B)	Decouple design to adapt open-sourced VLM for robotic control.
RoboMamba [112]	CLIP / SigLIP ViT-L	Mamba-2.8B/1.4B	PD (A), SFT (B)	Introduce the Mamba architecture to the VLA field.
DeeR-VLA [36]	CLIP ViT-L/14	MPT-1B / 7B	PD (A)	Propose dynamic early-exit to reduce inference overhead.
OpenVLA-OFT [44]	DINOv2 + SigLIP	LLaMA2-7B	PD (A)	Boost performance via OpenVLA-based fine-tuning.
PD-VLA [113]	CLIP ViT-L	Vicuna1.5-7B	PD (A)	Introduce parallel decoding manner for faster robot control.
MoLe-VLA [114]	DINOv2 + SigLIP	LLaMA2-7B	PD (A)	Reduce computation via dynamic LLM layer activation.
NORA [45]	-	Qwen2.5-VL-3B	AD (A)	Build efficient low-parameter model to boost performance.
FLashVLA [115]	DINOv2 + SigLIP	LLaMA	AD (A)	Propose retraining-free acceleration to improve VLA inference.
BitVLA [116]	SigLIP b1.58	BitNet b1.58 2B4T	PD (A), SFT (B)	Build ternary weight model to reduce deployment memory cost.
Spec-VLA [117]	DINOv2 + SigLIP	LLaMA2-7B	PD (A)	Propose speculative decoding to speed up without success drop.
CogVLA [118]	DINOv2 + SigLIP	LLaMA2-7B	PD (A)	Propose a cognition-aligned framework for efficient manipulation.

tuning, enabling multimodal fusion with limited modality-specific annotations. OE-VLA [104] leverages diverse multimodal data and a two-stage curriculum learning strategy, extending instructions beyond text to interleaved images, videos, and optical instructions.

Enhancing Reasoning Capabilities. To move VLA models from simple reactive control toward more advanced deliberative decision-making, enhancing reasoning ability is essential. As shown in “Enhancing Reasoning Capabilities” of Fig. 4, the LLM generates a chain-of-thought “Reasoning” process, which is then used as contextual information to produce the final “Action”. ECoT [121] generates a reasoning chain that sequentially combines high-level task planning with visually grounded features before outputting the final action. CoT-VLA [33] further introduces a visual chain-of-thought reasoning by predicting a subgoal observation that represents a planned state in pixel space. In contrast to reasoning-oriented perspectives, LoHoVLA [106] employs a “Hierarchical Closed-Loop Control” mechanism to ad-

dress planning errors, action failures, and external disturbances, thereby handling long-horizon tasks. ReFineVLA [105] adopts a “Selective Transfer Fine-Tuning” strategy with dual learning objectives to fine-tune only the upper layers, enabling the model to enhance multimodal understanding. These studies collectively underscore the central role of reasoning in enabling VLA models to achieve more reliable and generalizable action prediction.

Enhancing Generalization Capabilities. The generalization capability of a model refers to its ability to perform diverse tasks across different platforms and scenarios. As illustrated in “Enhancing Generalization Capabilities” of Fig. 4, UniAct [99] defines a “Universal Action Codebook” that abstracts heterogeneous robot actions into a unified representation. This unified encoding eliminates action-space heterogeneity and enables cross-embodiment learning and reasoning, thereby enhancing generalization. ReVLA [96] employs a reversible training strategy that gradually restores the vision encoder to its original pre-trained state,

mitigating catastrophic forgetting during fine-tuning and improving out-of-distribution visual generalization.

Another approach to enhancing generalization is to generate more robust and reliable actions, as depicted in Fig. 4. HybridVLA [34] integrates diffusion and autoregressive decoding within a unified model. It further employs a “Collaborative Action Ensemble” mechanism to adaptively fuse them, selecting the most suitable generation strategy for different tasks, thereby improving control robustness. Similarly, VOTE [109] introduces an adaptive “Ensemble Voting” strategy that groups past action predictions by similarity to the current one. It averages the majority set to produce more robust actions, balancing responsiveness and stability.

Enabling the model to better understand physical dynamics can promote the generation of more reliable actions. WorldVLA [38] integrates an action model and a world model within a unified autoregressive framework for capturing physical dynamics. Likewise, UnifiedVLA [107] introduces a world model as a post-training task, enabling it to learn causal dynamics of the physical world from large-scale unlabeled videos. UP-VLA [101] proposes a training framework that implicitly learns the physical dynamics of the world by predicting the next-frame image during pre-training. Consequently, these models generate more reliable actions that better adhere to physical laws.

3.1.3 Inference Efficiency Optimization

Large VLM-based VLA models often lead to slow inference speed, which conflicts with the high control-frequency requirements in robotic manipulation. Therefore, many works focus on inference efficiency optimization, mainly improving efficiency from the following three perspectives:

Architectural Optimization. These works aim to improve efficiency by adjusting the VLM architecture in VLA. As illustrated in “Architectural Optimization” of Fig. 4, dynamic inference can be achieved through layer skipping or early exiting, avoiding redundant computation. MoLe-VLA [114] introduces a “Spatio-Temporal Aware Router (STAR)” over the joint projection space of visual and language inputs, dynamically computing routing weights to select the most relevant LLM layers. This mechanism efficiently activates critical layers while skipping redundant ones, enabling fast and semantically sensitive inference. DeeR-VLA [36] enables dynamic inference by adding multiple early exits within the LLM. Adjacent exits compare output consistency to determine whether inference can be terminated early, avoiding unnecessary deep-layer computation. CogVLA [118] proposes a cognition-aligned routing framework that introduces instruction-driven sparsification across the perception-language-action pipeline. By progressively compressing visual tokens, pruning instruction-irrelevant grounded tokens, and coupling them with efficient V-L-A attention, CogVLA achieves both reduced inference latency and advanced success rates on robotic manipulation tasks. Unlike dynamic inference, RoboMamba [112] introduces Mamba-based LLMs [122] to the VLA domain, achieving over 3 \times faster inference than Transformer-based [91] models due to Mamba’s linear complexity with sequence length.

Parameter Optimization. Reducing model size directly lowers deployment difficulty. As shown in Fig. 4 under “Pa-

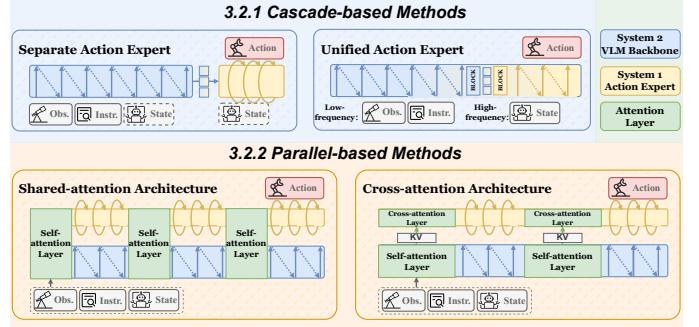


Fig. 5: Comparison of representative paradigms in monolithic dual-system models. Sec. 3.2.1 introduces cascade-based methods, where the VLM backbone is cascaded with the action expert to forward its output features in a single pass to the action expert. Sec. 3.2.2 introduces parallel-based methods, where the VLM backbone and the action expert operate in parallel and interact in various ways.

rameter Optimization”, model compression directly reduces the parameter count. BitVLA [116] introduces the first “1-bit” weight VLA model, using distillation-aware training to quantize the vision encoder and adopting the “1-bit” weight BitNet b1.58 [123] as its LLM. On the other hand, NORA [45] pairs a compact, high-quality VLM with a FAST+ [124] tokenizer that compresses high-dimensional actions into short token sequences. It demonstrates that a well-designed small model can outperform larger models.

Inference Acceleration. This line of work seeks to enhance inference efficiency by accelerating the process of action decoding during inference. The most notable aspect is that most models replace autoregressive decoding with parallel decoding. Fig. 4 under “Inference Acceleration” shows that decoding speed can be improved by modifying the LLM decoding process or using a specialized action head. In this way, a complete action sequence can be generated in a single forward pass. RoboFlamingo [111] separates perception and policy, using an independent MLP action head to convert VLM outputs into actions. This enables the generation of stacked actions in one pass. OpenVLA-OFT [44] replaces the causal attention mask with a bidirectional one, enabling empty action embeddings to be filled in one pass and converted to continuous actions via an MLP. PD-VLA [113] reformulates autoregressive decoding as a non-linear equation system solved via fixed-point iteration. Bidirectional attention updates action tokens until convergence, thereby enabling parallel decoding.

Different from parallel decoding, Spec-VLA [117] explores speculative decoding in VLA models. With Relaxed Acceptance to increase the verification model’s acceptance rate for draft model predictions, it achieves a 1.42 \times speedup over the baseline. Beyond decoding, FlashVLA [115] addresses redundancy by using FlashTrigger to evaluate the stability of actions and the environment and decide whether to skip the current inference, significantly reducing latency.

3.2 Dual-system Models

The need to simultaneously handle the demands of deep reasoning and real-time action generation has prompted the

development of monolithic dual-system VLA architectures. These models adopt two distinct systems: a larger and slower System 2 for high-level reasoning and planning, and a smaller and faster System 1 for low-level action generation. This separation enables a faster response time and smoother real-world deployment. It delivers both solid strategic planning and high-speed control for robotic manipulation. The separated low-level control module is also more adaptable to task-specific optimization. The key distinction between this approach and single-system models lies in the introduction of an action expert to explicitly decouple action generation. In contrast to hierarchical models, its major difference is that it does not produce an interpretable intermediate output. Tab. 2 summarizes representative dual-system methods. Fig. 5 illustrates two representative architectures: cascade-based methods and parallel-based methods.

3.2.1 Cascade-based Methods

As shown in the upper part of Fig. 5, cascade-based dual-system VLA models separate high-level semantic reasoning from low-level real-time control in a serial manner. The high-level system (System 2) typically employs VLMs to process multimodal input, perform semantic grounding, and generate action plans. These plans are encoded as latent cognitive representations rather than being executed directly. The low-level system (System 1) then decodes these representations into executable robot actions at a higher frequency, significantly improving run-time efficiency and facilitating real-world deployment. Fig. 5 shows two architectures, differing in how the action expert is designed.

To decouple the capabilities of “cognition” and “action” of the model, many methods select a separate model to serve as the action expert. As shown in the “Separate Action Expert” part at the top-left of Fig. 5, the VLM backbone transmits features to the action expert to convey information derived from visual, textual, and robot-specific state inputs. For example, CogACT [128] introduces the diffusion transformer (DiT) [144] as an action model. In particular, it also proposes an adaptive action ensemble algorithm for smoother and more efficient movement trajectories. GR007 N1 [32] also adopts a dual-system architecture in which DiT is introduced as the low-level action model. Both systems are tightly coupled and jointly trained end-to-end, enabling much faster policy steps. Some models adopt a similar architecture, but do not employ a DiT. For example, DP-VLA [125] uses a Behavioral Cloning transformer [145] instead.

Many models have made variations based on this. The HiRT model [129] uses VLM to operate at a lower frequency. This process extracts features for long-term understanding of the scene. Then a lightweight visual action strategy is deployed at a higher frequency. This design enables more efficient robotic manipulation and allows the system to keep up with changes in the real-world environment during deployment. TriVLA [133] is a three-module system incorporating a video generation model to predict future frames, a VLM to interpret instructions, and a diffusion action expert. GF-VLA [134] extracts information-based hand-object and object-object scene graphs from human demonstrations, then fuses them with an LLM to generate interpretable behavior trees and low-level Cartesian actions for dual-arm control. RationalVLA [135] couples a VLM with a diffusion

policy through a learnable latent interface. The VLM emits a token serving as the condition of the controller for action generation or a token to refuse infeasible commands. VQ-VLA [136] adopts a convolutional residual VQ-VAE [146], which is pretrained with action sequence, to take the place of the binning method of OpenVLA [26]. The model shows linear performance gains from more simulated data and has less sim-to-real gap. Some models integrate the action expert into the VLA backbone, Fast-in-Slow [40] is a typical example. As shown in the “Unified Action Expert” part at the top-right of Fig. 5, the action expert utilizes the final transformer blocks of the VLM backbone. They run at different frequencies, enabling seamless coordination between the two systems within a single pretrained model.

3.2.2 Parallel-based Methods

As shown in the lower part of Fig. 5, parallel-based dual-system VLA designs an action expert operating in parallel with the VLM backbone. The two components interact to exchange information during inference. Based on the choice of action expert and the interaction mechanism, two categories in Fig. 5 are defined.

As shown in the “Shared-attention Architecture” part at the bottom-left of Fig. 5, this architecture draws inspiration from the mix-of-experts (MoE) framework [147]. They leverage token interactions within self-attention layers, which are shared by the VLM backbone and the action expert to separate high-level reasoning from low-level execution. Information derived from visual and textual inputs interacts with noise and robot-specific inputs within the shared self-attention layers. This design facilitates task-specific optimization. For example, this design allows for the incorporation of a true MoE or other specially designed architectures within the action expert. A typical example is π_0 [29]. Its backbone weights are initialized from a pretrained VLM. To handle robot-specific inputs and action generation, a second set of independent weights, the flow-matching-based action expert, is introduced and trained from scratch. ForceVLA [138] uses π_0 [29] as the base model, the FVLMoE module with MoE is used to introduce the force modality into VLA. OneTwoVLA [139] is based on π_0 [29] and can switch between two modes: explicitly reasoning and generating actions based on the most recent reasoning. This architecture makes it easier for the two systems to operate asynchronously and further improves efficiency.

Many innovations are built upon this. $\pi_{0.5}$ [30] builds on π_0 [29] by introducing an additional step. The VLA module integrates visual information to convert high-level prompts into more fine-grained subtask predictions, which are then processed by the dual-system architecture of π_0 [29]. $\pi_{0.5}\text{-KI}$ [37] builds on $\pi_{0.5}$ [30]. During training, it prevents gradients from the action expert from flowing into the VLM backbone during training to preserve the VLM’s knowledge advantage. $\pi_0\text{-FAST}$ [124] introduces a DCT-driven action tokenization approach that facilitates efficient autoregressive training of VLA models. GraspVLA [143] proposes a dual-system model via Progressive Action Generation (PAG), which unifies autoregressive perception tasks and flow-matching-based action generation into a Chain-of-Thought process. This design enables joint training on synthetic and Internet data, achieving direct sim-to-real transfer

TABLE 2: Dual-system VLA models. The “System 2 Backbone” column lists the VLM backbone used as the System 2 component in dual-system methods. The “System 1 Learning” column lists the learning methods used by the action experts as System 1. “Diff.” denotes diffusion-based learning, “FM” denotes flow-matching, “MSE” denotes mean squared error, “BCE” denotes binary cross-entropy, and “AR” denotes autoregressive learning.

Model	System 2 Backbone	System 1 Learning	Contribution
Cascade-based			
DP-VLA [125]	OpenVLA	Regression	Propose a dual-system architecture for robot manipulation with efficiency and performance.
RoboDual [126]	OpenVLA	Diff.	Combine a VLA-based generalist for reasoning and a DiT specialist for control.
LCB [127]	LLaVA	Diff.	Leverage an added special token to encode VLM reasoning and act as conditions for policy.
GR00T N1 [32]	Eagle-2	FM	Combine a VLM and DiT for humanoid robots manipulation.
CogACT [128]	OpenVLA	Diff.	Propose an action ensemble algorithm to integrate the action diffusion process into VLA.
HiRT [129]	InstructBLIP	Regression	Propose a dual-system model with System 2 running at a lower frequency.
Fast-in-Slow [40]	Prismatic	Diff., AR	Propose a unified dual-system model that embeds fast execution within a VLM-based reasoner.
OpenHelix [58]	LLaVA	Diff.	Conduct auxiliary training on the token bridging VLM and policy.
ChatVLA [130]	Qwen2-VL	Diff.	Unify vision-language-action via MoE-shared attention with separate perception/control FFNs.
ChatVLA-2 [131]	Qwen2-VL	Diff.	Enable open-world robotic reasoning via dynamic MoE routing and Reasoning-Following MLP.
Diffusion-VLA [132]	Qwen2-VL	Diff.	Merge Qwen2-VL reasoning with diffusion actions via FiLM-modulated reasoning injection.
TriVLA [133]	Eagle-2	Diff.	Introduce a world-dynamics perception module as system 3 to complement static perception.
GF-VLA [134]	LLaMA 2	Regression	Enable interpretable bimanual manipulation via information-theoretic graphs from human videos.
RationalVLA [135]	LLaVA-v1.5	Diff.	Introduce a learnable latent interface to enable instruction rejection for robust manipulation.
VQ-VLA [136]	OpenVLA	VQ-VAE	Develop a vector quantization-based action tokenizer for efficient and smoother control.
TinyVLA [137]	LLaVA	Diff.	Demonstrate that high-performance VLAs require no large-scale robotic pretraining.
Parallel-based			
π_0 [29]	PaliGemma	FM	Combine a pre-trained Vision-Language Model with a Flow Matching-based Action Expert.
π_0 -FAST [124]	π_0	AR	Propose a DCT-based action tokenization enabling efficient autoregressive VLA training.
$\pi_{0.5}$ [30]	PaliGemma	FM	Convert high-level prompts into more fine-grained subtask predictions before feeding into π_0
$\pi_{0.5}$ -KI [37]	PaliGemma	FM	Prevent gradients from the action expert from flowing into the VLM backbone during training.
ForceVLA [138]	π_0	Diff.	Treat force sensing as a first-class modality via MoE, improving contact-rich manipulation.
SmolVLA [31]	SmolVLM-2	FM	Propose a lightweight VLA with frozen SmolVLM-2 and flow-matching transformer.
OneTwoVLA [139]	π_0	FM	Integrate acting/reasoning in shared VLA backbone processing multi-view inputs.
Tactile-VLA [140]	π_0	FM	Integrate tactile sensing to enable force-aware, generalizable contact-rich manipulation.
GR-3 [141]	Qwen2.5-VL	FM	Combine VL data and few-shot trajectories for robust manipulation in long-horizon or unseen tasks.
villa-X [142]	PaliGemma	FM	Integrate proprioceptively grounded latent actions and robot actions in a joint diffusion process.
GraspVLA [143]	InternLM2	FM	Enable sim-to-real and open-vocabulary grasping with synthetic data.

and open-vocabulary grasping. villa-X [142] grounds latent actions in robot states and jointly models latent and robot actions via joint diffusion for structured vision-action integration. Tactile-VLA [140] integrates tactile sensing into VLA models with hybrid position-force control. As shown in the “Cross-attention Architecture” part at the bottom-right of Fig. 5, this architecture feeds the visual, textual, and state inputs into the VLM, where the self-attention layers generate key-value and passed key-value to the action expert’s cross-attention layers. For instance, SmolVLA [31] adopts this architecture. It improves efficiency by leveraging a lightweight and frozen VLM backbone and training only a downstream Flow Matching Transformer as the action expert. Another example is GR-3 [141], which unifies vision-language understanding with robot trajectory learning using flow-matching.

4 HIERARCHICAL MODELS

Hierarchical modelling serves as a foundational paradigm in large VLM-based VLA models, particularly in scenarios where long-horizon reasoning, spatial abstraction, or action decomposition is required. These models are typically composed of a high-level planner and a low-level policy, as illustrated in Fig. 3 right. The planner receives instructions and observations, transforming them into interpretable intermediate representations. The policy then accepts these

representations and generates action sequences or codes that robots can directly execute. A summary of hierarchical models is provided in Tab. 3.

Importantly, the planner and policy in hierarchical models can operate independently, without either module being strictly dependent on the other. This modularity enables flexible combinations: many works focus solely on designing the planner component, leveraging existing off-the-shelf policies for execution. Therefore, we divide hierarchical models into two categories: Planner-Only (Sec. 4.1) and Planner+Policy (Sec. 4.2). Unlike dual-system VLA models, which also involve multiple modules, the intermediate representations in hierarchical models are explicitly interpretable to humans. Based on their nature, each category can be further divided into subtask-, keypoint-, and program-based methods. A comparative analysis of monolithic and hierarchical models is presented in Sec. 4.3.

4.1 Planner-Only

4.1.1 Program-based Methods

In this approach, planners generate intermediate programs for robotic manipulation, which fall into two categories: robot-executable programs and auxiliary programs. Robot-executable programs are built on robot libraries and can be directly executed to control the robot. For instance, Chain-of-Modality [151] employs a multimodal prompting strat-

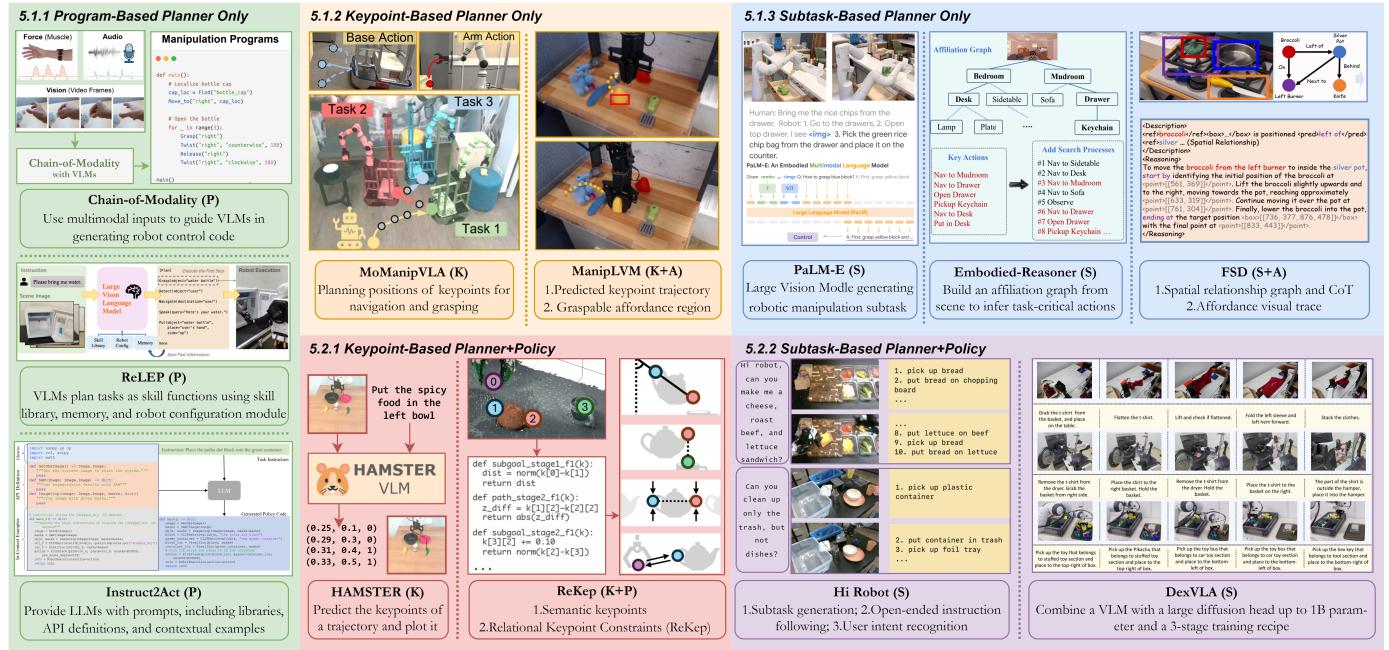


Fig. 6: A diagram showing the hierarchical models in this survey. These models are divided into two categories according to their constitutions, i.e., Planner-Only (Sec. 4.1) and Planner+Policy (Sec. 4.2). Based on the intermediate representation type, one category can be further divided into subtask-based (S), keypoint-based (K), and program-based (P). "(A)" denotes methods combining affordance for auxiliary purposes.

egy, where the VLM is engaged in a multi-turn conversation across different modalities and ultimately generates a robot-executable Python program to reproduce the task. Similarly, Instruct2Act [159] produces Python code that invokes APIs to control robot actions. In contrast, auxiliary programs support the policy in task understanding but cannot be executed directly. ROVI [152] exemplifies this category by generating auxiliary programs to describe the potential action and resolve the actual execution through translational and rotational costs. Likewise, ReLEP [153] uses a VLM with a memory bank to decompose tasks into basic skills from a skill library. It produces plans in the form of auxiliary programs that enable strong long-horizon performance.

4.1.2 Keypoint-based Methods

Keypoint-based planners use VLMs to predict salient points in an observation, typically corresponding to interactive regions that the gripper should reach (e.g., the handle of a drawer). Several methods tackle robotic manipulation by predicting waypoints. MoManipVLA [148] generates a key waypoint at each step through a VLA model, which is subsequently refined into executable actions via a bi-level trajectory optimization framework. Other approaches emphasize affordance-driven keypoints. RoboPoint [149] interprets natural language instructions to generate visual keypoints that specify precise manipulation targets. Besides, ManipLVM-R1 [46] trains a VLM with Group Relative Policy Optimization (GRPO) [173] to predict both the affordance area for grasping and the trajectory of the target object, thereby yielding a more generalizable planner. Similarly, RoboBrain [155] integrates task planning, affordance perception, and trajectory estimation by combining LLaVA [21] for high-level planning with an A-LoRA module to identify

interactable regions and a T-LoRA module to predict trajectory waypoints. This indicates that these models span both affordance and waypoints prediction, forming a hybrid representation. Moreover, some works explore distinctive forms of keypoints. RoVI [152], for instance, employs a sketch-based interface, extracting starting points, waypoints, and endpoints from hand-drawn annotations using YOLOv8, which are then applied as trajectory constraints.

4.1.3 Subtask-based Methods

In this approach, the planner is typically a large VLM that receives high-level implicit instructions (e.g., clean up the table) along with observations, and decomposes them into step-by-step textual commands. Since these models produce interpretable intermediate instructions rather than executable actions, a low-level control policy is still required in practical deployment. Early efforts such as PaLM-E [85], a LLaVA-style VLM trained with robotic manipulation data, demonstrate the feasibility of unifying general VQA capabilities with robot command generation. Building upon this direction, Embodied-Reasoner [47] introduces Observation–Thought–Action trajectories to support spatial analysis, reflection, and verification during step-wise planning. Reinforced Planning [150] retains this subtask-decomposition setup but boosts generalization through a two-stage pipeline of SFT followed by GRPO-based reinforcement fine-tuning. Meanwhile, some systems decouple perception from reasoning: Embodied-R [174] combines a large VLM for perception with a small LM for reasoning for embodied spatial reasoning on video, and its reasoning traces can be leveraged to derive stepwise subtask proposals for manipulation. In contrast, ViLA [154] leverages GPT-4V as an external planner, prompting it to generate candidate

TABLE 3: Hierarchical VLA models. The “Type” column denotes the output type of the planner, where “K” represents Keypoint, “S” represents Subtask, and “P” represents Program. The “Learning” column specifies the learning method adopted by the model, where “SFT” refers to Supervised Fine-Tuning, “RL” denotes Reinforcement Learning, “IM” indicates Imitation Learning, and “API” is a special case referring to the invocation of pre-existing models.

Model	Type	Backbone	Learning	Contribution
Planner-Only				
MoManipVLA [148]	K	OpenVLA-7B	IM	Leverage VLA models to predict waypoints and optimize full-body trajectories.
ManipLVM-R1 [46]	K	Qwen2.5-VL-3B	RL	GRPO tuning for affordance and trajectory, robust performance in OOD situations.
PaLM-E [85]	S	PaLM	SFT	Train a VLM capable of general VQA and robot manipulation instruction generation.
Emb-Reas [47]	S	Qwen2-VL-7B	SFT	Construct a VLM and open-sourced dataset with planning, reasoning, and reflection.
RoboPoint [149]	K	Vicuna-v1.5-13B	SFT	Finetune VLM for spatial affordance prediction in the form of keypoints.
Reinforced [150]	S	Qwen2.5-VL-7B	SFT, RL	Conduct GRPO on a finetuned model, bringing better generalization to unseen.
CoM [151]	P	Gemini 1.5 Pro	API	Sequential multimodal prompting to extract force-aware manipulation from demos.
RoVI [152]	K, P	GPT-4o / LLaVA-13B	SFT	Visual sketch-based instruction and hierarchical pipeline for precise manipulation.
ReLEP [153]	P	LLaVA-1.6-7B	SFT	A planning framework with implicit logical inference and hallucination mitigation.
ViLa [154]	S	GPT-4V	API	A VLM planner integrating perception and reasoning without affordance models.
RoboBrain [155]	K, S	LLaVA	SFT	Provide a hierarchical VLA focus on planning, affordance, and trajectory.
Planner+Policy				
HAMSTER [48]	K	VILA-1.5-13B	SFT, IM	Propose an out-of-the-box way for trajectory prediction to assist the low-level policy.
HiRobot [156]	S	PaliGemma-3B, π_0	SFT, IM	A hierarchical VLA with high explainability and capacity for complex tasks.
Agentic Robot [157]	S	GPT-4o	SFT	A closed-loop hierarchical pipeline where a VLM is attached to a completion judge.
DexVLA [158]	S	Qwen2-VL	SFT, IM	Combine a VLM with a large diffusion head up to 1B and a 3-stage training recipe.
Instruct2Act [159]	P	ChatGPT	API	Generate programs that call APIs for mapping from instructions to actions.
RoboMatrix [160]	S	Vicuna 1.5	SFT	VLA with modular scheduling layer, skill layer, and hardware layer.
PointVLA [161]	S	Qwen2-VL	SFT	Attach VLA with a point cloud encoder and injector to equip spatial perception.
A_0 [49]	K	Qwen2.5-7B	SFT, API	Hierarchical affordance-aware diffusion with embody-agnostic keypoint prediction.
FSD [162]	S	CLIP, Vicuna	SFT	Propose the generation of visual aids via SrCoT for zero-shot manipulation.
RoBridge [163]	S	GPT-4o	IM, RL	Bridge VLM cognition with RL execution via invariant operable representation.
Robocerebra [164]	S	GPT-4o, Qwen2.5-VL	SFT	A novel benchmark and hierarchical framework for long-horizon evaluation.
DexGraspVLA [165]	S	Qwen-VL	API, IM	Combine a VLM as a high-level planner with a low-level diffusion-based policy.
RT-H [28]	S	PaLI-X 55B	SFT, IM	An action hierarchy architecture using language motion as a middle representation.
ReKep [50]	K, P	GPT-4o	API	Training-free trajectory generation by keypoint constraints for manipulation.
VoxPoser [166]	P	GPT-4	API	Propose language-guided 3D value maps with zero-shot generalization capabilities.
SkillDiffuser [167]	S	Transformer	IM	A hierarchical VLA with high-level model and low-level model.
RT-Affordance [168]	K	PaLM-E 2	SFT, IM	Use visual affordances as intermediate features for web-robot knowledge transfer.
HiBaNAC [169]	S	PaLM2	SFT	Propose an asynchronous model that mimics the hierarchical structure of the brain.
LLARVA [170]	K	Llama 2 7B	SFT, IM	A vision-action instruction tuning paradigm and a large instruction tuning dataset.
MALMM [171]	S, P	GPT-4-Turbo	API	Three-agent system combining planner, supervisor, and coder without post-training.
VLA-Touch [172]	S	GPT-4o	IM	Integrate tactile sensing into VLA control via diffusion-based imitation learning.

task plans from textual and visual inputs and then executing only the first command, discarding the remainder.

4.2 Planner+Policy

4.2.1 Keypoint-based Methods

This type of model generally uses a large VLM to ground subgoals as spatial primitives, such as discrete keypoints or a 2D path on the image. The low-level policy then consumes these primitives to predict continuous trajectories and control. For instance, HAMSTER [48] first predicts trajectory keypoints from instructions and observations. Then it links these keypoints into an ordered path with a gradient color, and overlays this guidance for policy execution. Besides, ReKep [50] utilizes DINOv2 [93] and SAM [175] to produce keypoint proposals, then uses GPT-4o [176] to turn those keypoints into cost functions. These functions are subsequently solved into waypoints and actions by an optimizer. Moreover, A_0 [49] adopts an affordance-aware hierarchy: the planner predicts contact points and post-contact motion as an embodiment-agnostic affordance representation, which the action policy then converts into control.

4.2.2 Subtask-based Methods

Subtask-based hierarchical models bridge the planner and policy with instructions. The planner plays the same role as a subtask-based planner-only hierarchical model, and a low-level policy is appended to generate action sequences.

A representative example is HiRobot [156], the planner of which accepts open-ended user instructions and then decomposes them into atomic commands for the policy. Similarly, DexVLA [158] features a VLM planner and a diffusion-based action policy. The policy is conditioned on the subtask tokens given by the VLM planner, excelling in complex and long-horizon robotic manipulation tasks. Built on DexVLA, PointVLA [161] enhances spatial perception by incorporating a point cloud encoder and injector into the policy, enabling the model to follow the planner’s instructions in geometrically complex scenes. RoBridge [163] prompts the planner to generate text instructions of primitive actions and form an invariant operable representation for the policy to execute. Besides, SkillDiffuser [167] decomposes complex tasks into subtasks through a high-level model that predicts a set of skills, while a low-level diffusion policy realizes concrete actions. Analogously, RoboMatrix [160] organizes execution into a three-layer hierarchy: a

modular scheduling layer generates subtask sequences, the skill layer encodes and selects reusable behaviors, and the hardware layer implements robot control. HiBerNAC [169] proposes an asynchronous, hierarchical framework. In this framework, the multi-agent neural structure first decomposes high-level instructions into structured subtasks. Then, an asynchronous pipeline manages these subtasks and coordinates the reactive VLA to execute the final low-level actions. Moreover, MALMM [171] includes a planner, a supervisor, and a coder. The planner generates subtasks for the coder, and the supervisor coordinates transitions between modules. The coder serves as the policy and converts plans into executable robot code, including actions and positions.

4.3 Comparison: Monolithic vs. Hierarchical

The difference between monolithic and hierarchical architectures in large VLM-based VLA models for robotic manipulation lies primarily in how they map visual inputs and linguistic instructions into actions, either through a unified or modular approach. Monolithic models emphasize a single, integrated pipeline that jointly optimizes perception, reasoning, and control to directly translate high-level multimodal semantics into low-level actions. This design enables holistic and tightly coupled learning in robotic manipulation. Conversely, hierarchical architectures adopt a multi-stage design that explicitly separates high-level planning from low-level policy execution, promoting modularity and interpretability. This system decomposition allows components to be independently designed, trained, or replaced, enhancing flexibility and easing the integration of domain knowledge or adaptation to new robotic manipulation tasks.

Another core distinction lies in the inherent nature of the intermediate processing. Monolithic models, while potentially embedding intricate internal reasoning, do so within latent spaces that are often opaque to external inspection. This implicit strategy allows them to fully exploit the representational capacity of large models, potentially discovering efficient task decompositions not explicitly designed by humans. Hierarchical systems, on the other hand, explicitly commit to generating explicit, human-understandable intermediate outputs. This makes them particularly advantageous in robotic manipulation scenarios that demand explainability, detailed task monitoring, or compatibility with traditional robotics pipelines, where high-level plans can be independently validated or modified.

Despite these differences, both approaches contribute unique strengths to the evolving landscape of VLA for robotic manipulation. Monolithic models highlight the power of unified learning and minimal manual decomposition, offering a streamlined route to generalization across diverse tasks. Hierarchical frameworks, by explicitly layering cognition and control, provide greater transparency and enhanced modular flexibility, which can be critical for complex multi-stage tasks or safety-critical deployments. Together, these paradigms illustrate complementary strategies for bridging vision-language understanding and embodied action, each offering valuable insights into the design of next-generation intelligent robotic manipulation systems.

5 OTHER ADVANCED FIELD

Beyond core VLA architectures, we highlight four directions for robustness, efficiency, and long-horizon planning: RL-based optimization, training-free improvements, learning from human videos, and world model-based VLA. Tab. 4 summarizes representative methods.

5.1 Reinforcement Learning-based Methods

Reinforcement learning (RL) plays an important role in enhancing VLA’s generalization ability [177] and task completion rate [178]. It can be broadly categorized as online, which optimizes policies through real-time interaction, and offline, which learns from pre-collected trajectories. Most VLA methods are built upon traditional RL algorithms. We list several approaches in Tab. 5.

Unlike RL in LLMs or VLMs [79], [179], [180] that usually finish generation in a few turns, VLA models typically involve long-horizon tasks with hundreds of steps in one trajectory, making rule-based reward functions prone to sparsity and instability [177], [181]. To address this, many works incorporate learned dense reward signals. VLA-RL [181] trains a Robotic Process Reward Model (RPRM) to predict the success likelihood of action sequences. ReWiND [39] models reward as progress toward the final goal state, assigning higher rewards to states visually closer to task completion. In addition to training a reward model, Grape [177] and TGRPO [182] leverage powerful VLMs [176], [183] via prompting to generate feedback-based reward signals.

Besides reward sparsity, online learning is sample-inefficient due to slow simulation. To address this, several methods adopt hybrid offline-online training. ReWiND integrates offline Implicit Q-Learning [184] with online Soft Actor-Critic [185], accelerating learning and enhancing safety in real-world deployment. HIL-SERL [186] is based on RLPD and introduces human-in-the-loop intervention in the training process. ConRFT [187] employs a two-phase training scheme: an offline phase using Cal-ConRFT (combining behavior cloning and Q-learning) to initialize policy, followed by online HIL-ConRFT, which balances supervised and RL losses with human-in-the-loop intervention.

Furthermore, some methods leverage RL as a data engine to enhance generalist robotic models. RLDG [188] is a prime example, which trains expert policies via HIL-SERL [186] to 100% task success, then distills them into foundation models via sft. Similarly, iRe-VLA [189] alternates between online RL (to collect new successful trajectories) and sft on the expanded dataset, progressively improving the model through iterative imitation and exploration.

5.2 Training-Free Methods

Training-free methods typically leverage modular and extensible designs to improve existing VLA architectures without training. This enables rapid prototyping, ablation studies, and targeted enhancements while preserving the model’s original capabilities and avoiding additional costs.

A range of training-free methods has been proposed to improve the efficiency of large VLM-based VLA models efficiency without retraining or architectural changes. A compelling example is FlashVLA [115], which employs a

TABLE 4: Representative VLA methods grouped into four advanced categories: Reinforcement Learning-based, Training-Free, Learning from Human Videos, and World Model-based approaches. The grouping follows the methodological analysis in Sec. 5.

Description	Method
Reinforcement Learning-based Methods	
Use RL to optimize robotic policies through interaction or pre-collected trajectories.	VLA-RL [181], ReWiND [39], Grape [177], TGRPO [182], HIL-SERL [186], ConRFT [187], RLDG [188], iRe-VLA [189]
Training-Free Methods	
Improve VLA models via architectural or computational optimizations without retraining.	FlashVLA [115], EfficientVLA [190], VLA-Cache [191], PD-VLA [113], SP-VLA [192], BAC [193], FAST [124], RTC [41]
Learning from Human Videos	
Leverage human videos to adapt robot policies, enabling cross-domain transfer.	Human-Robot Semantic Alignment [42], UniVLA [194], LAPA [195], VPDD [196], 3D-VLA [197], Humanoid-VLA [198]
World Model-based VLA	
Integrate predictive world models into VLA to model environment dynamics.	World-VLA [38], World4Omni [43], 3D-VLA [197], RIGVid [199], FoundationPose [200], V-JEPA 2-AC [201]

TABLE 5: Representative RL approaches for VLA. ✓ stands for online and ✗ stands for offline. “S” stands for sparse reward, “D” for dense reward, “RM” denotes a pre-trained reward model, “GPT” denotes a reward given by GPT, and “TC” denotes a reward function based on task completion.

Method	RL Algo.	Online	Reward Formulation
VLA-RL [181]	PPO	✓	TC (S) + RM (D)
RIPT-VLA [202]	LOOP	✓	TC (S)
RLVLA [178]	PPO	✓	TC (S) + Object Grasp (S)
TPO [177]	DPO	✓	TC (S) + GPT (S)
ConRFT [187]	-	✗ + ✓	TC (S)
TGRPO [182]	GRPO	✓	TC (S) + GPT (D)
iRe-VLA [189]	PPO	✓	TC (S)
RLDG [188]	HIL-SERL	✓	TC (S)
ReWiND [39]	IQL SAC	✗ + ✓	TC (S) + RM (D)
HIL-SERL [186]	RLPD	✗ + ✓	TC (S)

trigger mechanism that skips full decoding when action and visual cues remain stable, selectively reusing or pruning visual tokens. Similarly, EfficientVLA [190] prunes redundant language layers, filters task-relevant visual tokens, and caches intermediate features, while VLA-Cache [191] reuses cached key-value representations [91] of static tokens with task-relevance filtering and layer-adaptive reuse. Beyond these, SP-VLA [192] combines spatio-semantic token pruning with an action-aware scheduler that routes intuitive steps to a lightweight generator and complex ones to the full VLA. Likewise, PD-VLA [113] reformulates autoregressive decoding under action chunking as a parallel fixed-point iteration, enabling simultaneous token prediction. FAST [124] compresses action sequences via discrete cosine transform and byte-pair encoding to reduce redundancy in high-frequency control, achieving up to 5x faster training. Furthermore, RTC [41] optimizes control-time scheduling by monitoring task progress and adjusting control frequency, reducing unnecessary computation during inference.

5.3 Learning from Human Videos

Recent advancements in VLA models have introduced several novel learning paradigms that surpass conventional policy learning strategies. Among them, a notable trend emerges: leveraging human video data to guide robot policy learning, using structural similarities in human-object and robot-object interactions to align visual and temporal cues.

This method aims to narrow the embodiment gap by transferring task-relevant knowledge from rich human video data. It uses human videos to pre-train perception or adapt policies. For instance, Human-Robot Semantic Alignment [42] aligns vision encoders across domains using paired human-robot videos, while UniVLA [194] learns task-centric latent actions from unlabeled human and robot videos to unify policy planning. Similarly, LAPA [195] leverages VQ-VAE [146] quantized latent actions to pre-train on large-scale video-language pairs, enabling transfer from actionless human videos. Besides, VPDD [196] applies discrete diffusion over unified video tokens, facilitating cross-domain knowledge transfer through future dynamics prediction. 3D-VLA [197] integrates human-object interaction videos with robot demonstrations for richer 3D reasoning, and Humanoid-VLA [198] exploits pose-recovered motion trajectories from online videos to enhance motion diversity. Collectively, these methods demonstrate that human video data can substantially enrich robot policy learning, yielding robustness and versatility even with limited robot data.

5.4 World Model-based VLA

World models, characterized by their ability to learn compact latent representations of environment dynamics, have emerged as powerful tools for enabling predictive reasoning and long-horizon planning. Recently, the integration of world models into VLA systems has gained traction as a promising strategy to enhance action planning through explicit modeling of environment dynamics.

Rather than directly generating actions from current observations, these approaches simulate future states, enabling agents to anticipate the consequences of their actions and refine decision-making accordingly. For instance, WorldVLA [38] introduces an autoregressive action world model that jointly learns to predict visual outcomes and generate actions within a unified token-based architecture. This mutual reinforcement between the world model and the action model improves both visual imagination and action fidelity, supporting more robust long-horizon planning. World4Omni [43] employs a large-scale world model to produce subgoal images that depict intermediate task states. These generated visual cues are then used to guide a modular low-level policy, enabling zero-shot manipulation across diverse environments and robot embodiments. 3D-VLA [197] employs a generative world model to predict future goal images and point clouds conditioned on instructions, effectively simulating scene dynamics. RIGVid [199] uses a diffusion-based world model to generate candidate task videos, which are then filtered by a VLM for feasibility, with FoundationPose [200] extracting gripper poses for execution. V-JEPA 2-AC [201] builds a latent action-conditioned world model on top of a massive self-supervised video encoder trained on internet-scale data. During inference, it

performs model-based planning by simulating future latent states conditioned on goal images, bridging dense video understanding with zero-shot robotic manipulation.

6 CHARACTERISTICS OF VLA MODELS

6.1 Multimodal Fusion

Shared Embedding Space. Large VLM-based VLA models embed visual observations and linguistic instructions into a shared, semantically aligned latent space. Unlike traditional robotic manipulation models that modularize perception, instruction parsing, and control, Large VLM-based VLA models rely on pre-trained VLM backbones to generate joint embeddings for both modalities. This unified space facilitates tight semantic grounding between perception and command, improving action decoding fidelity and reducing inter-module semantic drift [8].

Multimodal Token-Level Integration. A core characteristic in many VLA models is discretizing continuous modalities such as vision, language, proprioception, and actions into token sequences processed by a single transformer [40], [107], [203]. This design enables token-level multimodal integration and fine-grained coordination: linguistic tokens like “grasp red cup” immediately direct attention to relevant visual tokens, while proprioceptive and action tokens guide execution. The interleaved temporal structure naturally captures cross-modal dependencies across the perception–action cycle. Compared to traditional methods that fuse modalities at pre-defined stages, this dynamic integration reduces latency and semantic fragmentation.

Comprehensive Modal Compatibility. Another notable characteristic of VLA models lies in their inherent ability to seamlessly accommodate diverse sensory modalities, such as depth (e.g., point clouds), tactile input, and ambient audio. This compatibility stems from the modality-agnostic semantic alignment inherited from pre-trained VLMs, which allows new sensory tokens to be seamlessly integrated without altering the core architecture or requiring full model retraining. For instance, PointVLA [161] integrates point cloud data into pre-trained VLA architectures to enrich spatial perception and generalization, without retraining the backbone. Similarly, 3D-VLA [197] aligns point cloud features with language embeddings to achieve a more precise understanding and reasoning over complex 3D environments. Likewise, tactile and auditory tokens improve responsiveness to environmental subtleties. In contrast, traditional robotic manipulation models require costly redesigns when adding new modalities, making VLA models more scalable and robust for real-world deployment.

6.2 Instruction Following

Semantic Instruction Grounding. Instruction following in large VLM-based VLA models transcends the brittle and fragmented pipelines of traditional robotic manipulation models. Specifically, VLA models harness the rich world knowledge of pre-trained VLMs to interpret and ground natural language instructions into actionable behaviors dynamically. Rather than reducing instructions to fixed templates, they enable fluid, context-sensitive comprehension. For instance, ChatVLA-2 [131] can interpret math problems

written on whiteboards and select the correct number cards. Despite lacking explicit training in mathematics, it demonstrates an ability to generalize from visual-linguistic priors.

Task Decomposition and Collaboration. Beyond direct instruction-to-action translation, instruction following often involves hierarchical task decomposition, especially for long-horizon tasks. Specifically, high-level commands are broken down into sub-goals, which are then executed by low-level controllers [156]. LoHoVLA [106] serves as a compelling example that illustrates how this paradigm enables effective instruction following. It generates intermediate subtasks in natural language from visual and linguistic inputs. Then these subtasks guide downstream action generation and ensure semantic alignment throughout execution. By continuously synchronizing subtask intentions with actions, the model maintains coherence in long-horizon tasks, supporting reliable goal completion.

Explicit Reasoning via Chain-of-Thought. A key fact of the instruction following capability of VLA models is the integration of chain-of-thought reasoning into the decision-making loop. In CoT-VLA [33], for instance, the model interposes latent visual goals—predicting future images—before generating corresponding action sequences, effectively embedding a sub-goal generation mechanism. This explicit reasoning ties instructions to anticipated visual outcomes. It helps mitigate shortsighted hallucinations frequently observed in monolithic systems. In addition, it allows iterative refinement of plans, thereby facilitating more reliable execution of complex, multi-step tasks.

6.3 Multi-Dimensional Generalization

Cross-Task Generalization. Most VLA models exhibit remarkable zero-shot and few-shot generalizations, representing a leap beyond the brittle performance of traditional robotic manipulation models that often require extensive retraining. A prime example is DexVLA [158], which couples a large VLM with a billion-parameter diffusion action expert and an embodied curriculum to transfer skills across embodiments and rapidly adapt to unseen tasks. Without task-specific tuning, it attains consistently high success on diverse manipulations and generalizes to novel objects and scenes. Under direct language prompting, it also completes complex long-horizon workflows and outperforms state-of-the-art baselines such as OpenVLA [26] and π_0 [29].

Cross-Domain Data Generalization. This capacity extends further: $\pi_{0.5}$ and similar “foundation VLA” models are co-trained on heterogeneous data—web text, simulation videos, and cross-robot embodiments—to acquire broad contextual and semantic understanding. As a result, they successfully deploy “out-of-the-box” in new home environments, executing multi-stage tasks like dishwashing with OOD success rates exceeding 90% [204]. This wide compatibility starkly contrasts with traditional pipelines, which typically collapse outside their training distribution.

Cross-Embodiment and Sim-to-Real Generalization. Cross-embodiment and sim-to-real transfer represent another dimension of generalization. Hierarchical VLA architectures, which learn high-level planners in VLM space and delegate low-level control to domain-specific decoders, enable leveraging off-domain data (e.g., hand-drawn sketches

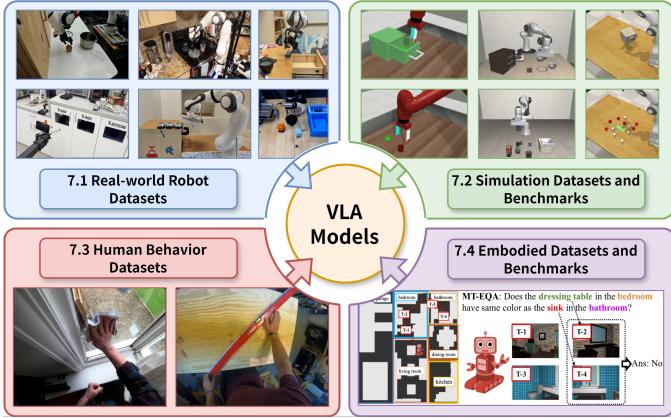


Fig. 7: Illustration of four dataset types that underpin large VLM-based VLA models for robotic manipulation.

or simulations in diverse setups) to generalize across varying dynamics and robot morphologies. One such model, HAMSTER [48], achieves a 20% success boost over OpenVLA across seven axes of generalization.

7 DATASETS AND BENCHMARKS

7.1 Real-world Robot Datasets

Real-world data, capturing environmental complexity such as dynamic lighting, background motion, and varied object properties, is vital for training and evaluating VLA models. Effective language grounding and semantic generalization rely on aligning natural language instructions, perceptual inputs, and precise action responses. To support this, a series of large-scale real-world datasets have emerged. BC-Z [205] provides expert demonstrations with language commands across 100 tasks; RT-1 [83] and its kitchen variant extend this to over 700 daily activities, while RT-2 [27] incorporates web-scale vision-language data for broader vocabulary and instruction generalization. RoboFlamingo [111] exemplifies the transfer of pretrained multimodal priors into real-world control. BridgeData V2 [206] and RH20T [207] further enhance cross-domain generalization through extensive, language-annotated demonstrations, with RH20T enabling one-shot learning across 147 tasks. DROID [208] expands this landscape with teleoperated demonstrations “in the wild,” offering diverse visual contexts over 564 tasks. Most notably, the OXE dataset [90] unifies over 1 million multimodal demonstrations across 22 robot platforms and 500+ skills, marking a significant step toward web-scale, cross-embodiment generalization. Although these efforts improve real-world applicability, the long-tail distribution of open-world objects, scenes, and skills remains underrepresented, underscoring the need for broader and more diverse real-world data to scale large VLM-based VLA models further.

7.2 Simulation Datasets and Benchmarks

To overcome real-world data limitations, physics-based simulators and virtual environments offer scalable, safe, and reproducible interaction data, aligning with the multimodal, large-scale training needs of large VLM-based VLA

models. They enable complex instruction following, multi-stage planning, and consistent language grounding with automatic evaluation. BEHAVIOR [209] supports multi-step semantic control in cluttered household settings; ALFRED [210] focuses on long-horizon tasks specified by egocentric language instructions; RLBench [211] allows fine-grained policy learning through RGB-D tabletop manipulation, and RLBench2 [212] extends it to bimanual manipulation with language-conditioned action prediction. Benchmarks such as Meta-World [213], Franka Kitchen [214], and LIBERO [215] further advance this direction by incorporating multi-skill scenarios, often enriched with language annotations and vision-language rewards to encourage generalization. CALVIN [216] enables multi-stage manipulation under unconstrained language instructions to support long-horizon behaviors. MIKASA-Robo [217] addresses memory-centric challenges by evaluating agents under partial observability in tabletop tasks. SIMPLER [218] bridges the sim-to-real gap through calibrated simulated environments aligned with common real-world robot configurations, reducing discrepancies in perception and control. High-fidelity platforms such as Habitat [219]–[221] and SAPIEN [222] support spatial reasoning and manipulation learning in realistic 3D environments, enabling scalable embodied VLA research. Beyond traditional simulation benchmarks, THE COLOSSEUM [223] and VLABench [224] extend simulation benchmarks by assessing robustness under distribution shifts and supporting universal language-conditioned manipulation, respectively. Despite limitations such as imperfect physics and visual artifacts, simulation remains crucial for efficient training, systematic evaluation, and robust real-world generalization.

7.3 Human Behavior Datasets

Large-scale human behavior datasets offer semantically rich and contextually diverse demonstrations that are well-suited for pretraining VLA models. Egocentric video corpora such as Ego4D [225], Ego-Exo4D [226], EgoPlan-Bench [227], and EgoVid-5 [228] capture diverse daily activities from a first-person perspective. In contrast, domain-specific datasets like EPIC-Kitchens [229] and COM-Kitchens [230] focus on fine-grained, temporally structured cooking tasks. These resources support learning of object recognition, action sequencing, and task decomposition. Reasoning-oriented datasets such as EgoVQA [231] and EgoTaskQA [232] further enhance cognitive grounding by introducing spatial, temporal, and causal reasoning over egocentric video. To address the lack of precise manipulation signals in general-purpose corpora like Ego4D, EgoDex [233] extends this line of work with 829 hours of egocentric video paired with dense 3D hand and finger tracking. Similarly, smaller manipulation-focused datasets such as DexCap [234], EgoMimic [235], and PH²D [236] offer fine-grained supervision for hand-object coordination and embodiment-aware policy learning. Complementing these human-centric datasets, R2R2R [237] offers a scalable pipeline that replaces teleoperation with smartphone scans and human videos to generate high-fidelity demonstrations. Together, these resources provide a rich spectrum of semantic grounding, embodiment alignment, and data scalability essential for training generalist VLA models.

7.4 Embodied Datasets and Benchmarks

As robots advance toward greater autonomy, understanding and planning tasks become essential. Leveraging strong language and world knowledge, VLMs empower VLA systems to plan over long horizons, parse complex instructions, and build cross-modal reasoning for semantic decision-making. In response, a growing set of benchmarks has emerged to evaluate embodied agents' planning and reasoning abilities. Early efforts such as EmbodiedQA [238] and IQUAD [239] established the integration of vision, language, and spatial navigation through question-answering in 3D environments. These have since evolved into more cognitively demanding tasks. MT-EQA [240] incorporates multi-target reasoning, MP3D-EQA [241] introduces 3D point cloud inputs for enhanced spatial grounding, EQA-MX [242] adds non-verbal cues for naturalistic multimodal interaction, and OpenEQA [243] expands the scope to open-ended questions involving functional and commonsense reasoning. Complementarily, LoTa-Bench [244] directly evaluates plan executability by deploying LLM-generated plans in simulation. Collectively, these benchmarks emphasize active perception, reasoning, and execution, offering rigorous protocols to assess high-level semantic planning in general-purpose VLA.

8 FUTURE DIRECTIONS

Datasets and Benchmarking. As VLA models advance, the reliance on simulated datasets reveals a significant reality gap. Synthetic scenes lack the visual complexity of real environments. By contrast, collecting real-world data is costly, limiting both diversity and scale. Meanwhile, existing benchmarks typically focus on short-horizon pick-and-place tasks and report simple success rates. These settings offer little insight into practical challenges such as long-term planning, mobile manipulation and multi-agent collaboration. Future datasets may combine large-scale real-world data acquisition with task suites that capture practical challenges. Evaluations should include richer metrics, such as subtask success, time efficiency, and robustness to disturbances. Such datasets and benchmarks would enable systematic progress on the broader capabilities of VLA systems.

Memory Mechanisms and Long-Term Planning. Effective real-world manipulation often requires planning over extended horizons and recalling past observations. However, most current VLAs rely on frame-by-frame reasoning, resulting in short-sighted behavior and a limited ability to leverage historical context. To address this, a promising direction is to design architectures that incorporate forward-looking planning and memory mechanisms with episodic awareness. By grounding high-level decisions in contextual memory, agents may move beyond reactive responses toward coherent, goal-driven action sequences, enabling more consistent and effective manipulation in complex scenarios.

3D and 4D Perception. Robotic manipulation inherently involves interacting with objects in a three-dimensional and temporally dynamic environment. However, existing VLA models operate primarily on static 2D visual inputs. This limits the agent's ability to reason about complex spatial and temporal information, including depth, affordances, object movement, and human actions. Moving beyond static 2D snapshots, 4D perception demands a model to understand

how 3D scenes evolve over time. The progression from 2D to 4D perception requires integrating depth or point cloud observations, as well as fusing multi-modal inputs into unified representations. It also involves embedding perception that is grounded in temporal context and replanning actions on the fly. Together, these 4D-aware abilities would enable more robust and adaptive manipulation in the real world.

Mobile Manipulation. Real-world tasks often demand the simultaneous execution of locomotion and manipulation, giving rise to the specialized domain of mobile manipulation. This task requires a synergistic integration of navigation capabilities and interactive skills with the environment, introducing tightly coupled demands on perception and control. Rather than treating locomotion and grasping as separate stages, future VLA models may benefit from learning integrated policies that adaptively prioritize locomotion and arm coordination, ultimately leading to more robust and flexible mobile manipulation.

Multi-Agent Cooperation. Many real-world collaborative tasks, such as bulky objects moving or joint tool use, require agents to negotiate intentions, adapt to teammates' actions, and jointly reason over multi-step goals. However, single-agent VLAs often falter when communication and role assignment become critical, highlighting the need for interaction-aware representations. Integrating emergent dialogue protocols with shared world models may enable agents to synchronize plans and flexibly allocate subtasks. Such cooperative capabilities promise to elevate robotic teams from loosely coordinated individuals to cohesive collaborators capable of addressing complex group objectives.

Lifelong Learning in Open-World. Robotic autonomy demands continual skill acquisition without catastrophic forgetting. However, most VLA models trained on static datasets struggle to handle unfamiliar objects and novel interaction modes. They also have difficulty incorporating new experiences. This reveals the limits of static training paradigms. Future VLA systems may benefit from mechanisms that incrementally accumulate knowledge through exploration and feedback in open-world. Incorporating memory structures that grow over time and enabling agents to self-organize experiences into reusable abstractions may also provide a foundation for long-term competence.

Model Efficiency. Though VLA models excel at understanding vision-language instructions and planning actions, they often suffer from prohibitive computational and memory costs. Deploying these models on resource-constrained robotic platforms raises concerns about latency, memory usage, and sustained operation. The key challenges lie in balancing model capacity with real-time inference requirements and preserving multimodal alignment to maintain accuracy during compression. Promising directions include task-aware dynamic token pruning, asynchronous inference for seamless transitions between action chunks, and hardware-friendly quantization schemes.

9 CONCLUSION

In conclusion, this survey provides the first principled synthesis of VLA models built on large VLMs. We trace the evolution of VLMs, analyze their integration into robotic

manipulation, and propose a two-part taxonomy: Monolithic (Single-system and Dual-system) and Hierarchical (Planner-Only and Planner+Policy). This taxonomy frames existing models and guides future VLA design. We also organize datasets, benchmarks, learning paradigms, and advanced methods into a coherent framework; identify defining characteristics such as open world generalization, multimodal reasoning, and instruction grounding in dynamic settings; and summarize trends toward modular planning and control, larger multimodal corpora, and closer links with reinforcement learning and world models. Looking ahead, priorities include adaptation across embodiments, scalable real-world deployment, tighter coupling between high-level reasoning and low-level execution, and the use of human demonstrations at internet scale. We aim for this survey to serve as a foundation for embodied AI that unifies perception, language, and action.

REFERENCES

- [1] Z. Xu, K. Wu, J. Wen, J. Li, N. Liu, Z. Che, and J. Tang, "A survey on robotics with foundation models: toward embodied ai," *arXiv:2402.02385*, 2024.
- [2] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv:2405.14093*, 2024.
- [3] W. Wu, H. He, Y. Wang, C. Duan, J. He, Z. Liu, Q. Li, and B. Zhou, "Metaurban: a simulation platform for embodied ai in urban spaces," in *ICLR*, 2025.
- [4] K. Zhang, P. Yun, J. Cen, J. Cai, D. Zhu, H. Yuan, C. Zhao, T. Feng, M. Y. Wang, Q. Chen *et al.*, "Generative artificial intelligence in robotic manipulation: a survey," *arXiv:2503.03464*, 2025.
- [5] Y. Liu, W. Chen, Y. Bai, X. Liang, G. Li, W. Gao, and L. Lin, "Aligning cyber space with physical world: a comprehensive survey on embodied ai," *arXiv:2407.06886*, 2024.
- [6] Y. Liu, X. Cao, T. Chen, Y. Jiang, J. You, M. Wu, X. Wang, M. Feng, Y. Jin, and J. Chen, "A survey of embodied ai in healthcare: techniques, applications, and opportunities," *INFORM FUSION*, vol. 119, p. 103033, 2025.
- [7] S. Nahavandi, R. Alizadehsani, D. Nahavandi, C. P. Lim, K. Kelly, and F. Bello, "Machine learning meets advanced robotic manipulation," *INFORM FUSION*, vol. 105, p. 102221, 2024.
- [8] R. Sapkota, Y. Cao, K. I. Roumeliotis, and M. Karkee, "Vision-language-action models: concepts, progress, applications and challenges," *arXiv:2505.04769*, 2025.
- [9] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, p. eaat8414, 2019.
- [10] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," in *RSS*, 2024, p. 92.
- [11] P.-L. Guhur, S. Chen, R. G. Pinel, M. Tapaswi, I. Laptev, and C. Schmid, "Instruction-driven history-aware policies for robotic manipulations," in *CoRL*, 2023, pp. 175–187.
- [12] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," in *CoRL*, 2024.
- [13] 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 *ICLR*, 2024, pp. 2475–2499.
- [14] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel, and E. Adelson, "Polytouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies," in *ICRA*, 2025.
- [15] V. Mengers and O. Brock, "No plan but everything under control: Robustly solving sequential tasks with dynamically composed gradient descent," in *ICRA*, 2025.
- [16] H. Li, Q. Lv, R. Shao, X. Deng, Y. Li, J. Hao, and L. Nie, "Star: Learning diverse robot skill abstractions through rotation-augmented vector quantization," in *ICML*, 2025.
- [17] G. Chen, L. Shen, R. Shao, X. Deng, and L. Nie, "Lion: empowering multimodal large language model with dual-level visual knowledge," in *CVPR*, 2024, pp. 26540–26550.
- [18] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024, pp. 26296–26306.
- [19] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, and S. Hoi, "Instructblip: towards general-purpose vision-language models with instruction tuning," in *NeurIPS*, 2023, pp. 49250–49267.
- [20] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: a versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv:2308.12966*, 2023.
- [21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023, pp. 34892–34916.
- [22] 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:2412.05271*, 2024.
- [23] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai, "Monkey: image resolution and text label are important things for large multi-modal models," in *CVPR*, 2024, pp. 26763–26773.
- [24] R. Zhang, R. Shao, G. Chen, M. Zhang, K. Zhou, W. Guan, and L. Nie, "Falcon: Resolving visual redundancy and fragmentation in high-resolution multimodal large language models via visual registers," in *ICCV*, 2025.
- [25] L. Shen, G. Chen, R. Shao, W. Guan, and L. Nie, "Mome: Mixture of multimodal experts for generalist multimodal large language models," in *NeurIPS*, 2024.
- [26] 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," in *CoRL*, 2024, pp. 2679–2713.
- [27] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, others, and K. Han, "Rt-2: vision-language-action models transfer web knowledge to robotic control," in *CoRL*, 2023, pp. 2165–2183.
- [28] 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 *RSS*, 2024.
- [29] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π₀: A vision-language-action flow model for general robot control," in *RSS*, 2025.
- [30] P. Intelligence, N. B. Kevin Black, K. D. James Darpinian, A. E. Danny Driess, C. F. Michael Equi, and e. a. Niccolo Fusai, "π_{0.5}: a vision language-action model with open-world generalization," *arXiv:2504.16054*, 2025.
- [31] 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:2506.01844*, 2025.
- [32] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: an open foundation model for generalist humanoid robots," *arXiv:2503.14734*, 2025.
- [33] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, T.-Y. Lin, G. Wetzstein, M.-Y. Liu, and D. Xiang, "Cot-vla: visual chain-of-thought reasoning for vision-language-action models," in *CVPR*, 2025, pp. 1702–1713.
- [34] 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:2503.10631*, 2025.
- [35] 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:2506.07961*, 2025.
- [36] 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 *NeurIPS*, 2024, pp. 56619–56643.
- [37] J. T. S. Danny Driess, L. Y. Brian Ichter, K. P. Adrian Li-Bell, H. W. Allen Z. Ren, L. X. S. Quan Vuong, and S. Levine, "Knowledge insulating vision-language-action models: train fast, run fast, generalize better," *arXiv:2505.23705*, 2025.
- [38] 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:2506.21539*, 2025.
- [39] J. Zhang, Y. Luo, A. Anwar, S. A. Sontakke, J. J. Lim, J. Thomason, E. Biyik, and J. Zhang, "Rewind: language-guided rewards teach

- robot policies without new demonstrations," *arXiv:2505.10911*, 2025.
- [40] 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:2506.01953*, 2025.
- [41] K. Black, M. Y. Galliker, and S. Levine, "Real-time execution of action chunking flow policies," *arXiv:2506.07339*, 2025.
- [42] J. Zhou, T. Ma, K. Y. Lin, Z. Wang, R. Qiu, and J. Liang, "Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation," in *CVPR*, 2025, pp. 22551–22561.
- [43] H. Chen, B. Wang, J. Guo, T. Zhang, Y. Hou, X. Huang, C. Tie, and L. Shao, "World4omni: a zero-shot framework from image generation world model to robotic manipulation," *arXiv:2506.23919*, 2025.
- [44] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: optimizing speed and success," in *RSS*, 2025.
- [45] C.-Y. Hung, Q. Sun, P. Hong, A. Zadeh, C. Li, U. Tan, N. Majumder, S. Poria *et al.*, "Nora: a small open-sourced generalist vision language action model for embodied tasks," *arXiv:2504.19854*, 2025.
- [46] Z. Song, G. Ouyang, M. Li, Y. Ji, C. Wang, Z. Xu, Z. Zhang, X. Zhang, Q. Jiang, Z. Chen *et al.*, "Maniplvm-r1: reinforcement learning for reasoning in embodied manipulation with large vision-language models," *arXiv:2505.16517*, 2025.
- [47] W. Zhang, M. Wang, G. Liu, X. Huixin, Y. Jiang, Y. Shen, G. Hou, Z. Zheng, H. Zhang, X. Li *et al.*, "Embodied-reasoner: synergizing visual search, reasoning, and action for embodied interactive tasks," *arXiv:2503.21696*, 2025.
- [48] Y. Li, Y. Deng, J. Zhang, J. Jang, M. Memmel, R. Yu, C. R. Garrett, F. Ramos, D. Fox, A. Li *et al.*, "Hamster: hierarchical action models for open-world robot manipulation," in *ICLR*, 2025.
- [49] 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:2504.12636*, 2025.
- [50] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekeep: spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," in *CoRL*, 2024.
- [51] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," in *IJCAI*, 2022.
- [52] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *TPAMI*, vol. 46, pp. 5625–5644, 2024.
- [53] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in *BigData*, 2023, pp. 2247–2256.
- [54] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges," *arXiv:2501.02189*, 2025.
- [55] D. Shu, H. Zhao, J. Hu, W. Liu, A. Payani, L. Cheng, and M. Du, "Large vision-language model alignment and misalignment: A survey through the lens of explainability," *arXiv:2501.01346*, 2025.
- [56] M. Song, X. Deng, Z. Zhou, J. Wei, W. Guan, and L. Nie, "A survey on diffusion policy for robotic manipulation: Taxonomy, analysis, and future directions," *Authorea Preprints*, 2025.
- [57] R. Wolf, Y. Shi, S. Liu, and R. Rayyes, "Diffusion models for robotic manipulation: A survey," *arXiv:2504.08438*, 2025.
- [58] 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:2505.03912*, 2025.
- [59] Y. Zheng, L. Yao, Y. Su, Y. Zhang, Y. Wang, S. Zhao, Y. Zhang, and L.-P. Chau, "A survey of embodied learning for object-centric robotic manipulation," *MIR*, vol. 22, pp. 588–626, 2025.
- [60] W. Li, B. Hu, R. Shao, L. Shen, and L. Nie, "Lion-fs: Fast & slow video-language thinker as online video assistant," in *CVPR*, 2025.
- [61] Z. Li, Y. Xie, R. Shao, G. Chen, D. Jiang, and L. Nie, "Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks," in *NeurIPS*, 2024.
- [62] —, "Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy," in *CVPR*, 2025.
- [63] Q. Ye, Z. Yu, R. Shao, X. Xie, P. Torr, and X. Cao, "Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios," in *ECCV*, 2024.
- [64] Q. Ye, Z. Yu, R. Shao, Y. Cui, X. Kang, X. Liu, P. Torr, and X. Cao, "Cat+: investigating and enhancing audio-visual understanding in large language models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [65] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: the convergence of autonomous driving and large vision-language models," in *CoRL*, 2025, pp. 4698–4726.
- [66] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding *et al.*, "Cogagent: a visual language model for gui agents," in *CVPR*, 2024, pp. 14281–14290.
- [67] G. Chen, X. Zhou, R. Shao, Y. Lyu, K. Zhou, S. Wang, W. Li, Y. Li, Z. Qi, and L. Nie, "Less is more: Empowering gui agent with context-aware simplification," in *ICCV*, 2025.
- [68] Y. Lyu, R. Shao, G. Chen, Y. Zhu, W. Guan, and L. Nie, "Puma: Layer-pruned language model for efficient unified multimodal retrieval with modality-adaptive learning," in *ACM MM*, 2025.
- [69] Y. Zhu, Y. Lyu, Z. Yu, R. Shao, K. Zhou, and L. Nie, "Emosym: A symbiotic framework for unified emotional understanding and generation via latent reasoning," in *ACM MM*, 2025.
- [70] B. Xie, R. Shao, G. Chen, K. Zhou, Y. Li, J. Liu, M. Zhang, and L. Nie, "Gui-explorer: Autonomous exploration and mining of transition-aware knowledge for gui agent," in *ACL*, 2025.
- [71] R. Shao, T. Wu, and Z. Liu, "Robust sequential deepfake detection," *International Journal of Computer Vision*, vol. 133, pp. 3278–3295, 2025.
- [72] R. Shao, T. Wu, J. Wu, L. Nie, and Z. Liu, "Detecting and grounding multi-modal media manipulation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 5556–5574, 2024.
- [73] R. Shao, T. Wu, and Z. Liu, "Detecting and grounding multi-modal media manipulation," in *CVPR*, 2023.
- [74] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *CVPR*, 2019.
- [75] R. Shao, T. Wu, L. Nie, and Z. Liu, "Deepfake-adapter: Dual-level adapter for deepfake detection," *International Journal of Computer Vision*, vol. 133, pp. 3613–3628, 2025.
- [76] J. Chen, D. Yuen, B. Xie, Y. Yang, G. Chen, Z. Wu, L. Yixing, X. Zhou, W. Liu, S. Wang, K. Zhou, R. Shao, L. Nie, Y. Wang, J. HAO, J. Wang, and K. Shao, "Spa-bench: A comprehensive benchmark for smartphone agent evaluation," in *ICLR*, 2025.
- [77] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li, "Llava-onevision: easy visual task transfer," *TMLR*, 2025.
- [78] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-v1 technical report," *arXiv:2502.13923*, 2025.
- [79] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, Y. Hu, and S. Lin, "Vision-r1: incentivizing reasoning capability in multimodal large language models," *arXiv:2503.06749*, 2025.
- [80] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: what and where pathways for robotic manipulation," in *CoRL*, 2022, pp. 894–906.
- [81] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, others, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.
- [82] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Atarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: rearranging the visual world for robotic manipulation," in *CoRL*, 2021, pp. 726–747.
- [83] A. Brohan, N. Brown, J. Carbalaj, 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 *RSS*, 2023.
- [84] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.
- [85] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: an embodied multimodal language model," in *ICML*, 2023, pp. 8469–8488.
- [86] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay *et al.*, "Pali-x: on scaling up a multilingual vision and language model," in *CVPR*, 2024, pp. 14432–14444.
- [87] 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 *ICLR*, 2022.
- [88] J. Wang, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, B. Ge, and S. Zhang, "Large language models for robotics: op-

- portunities, challenges, and perspectives," *JAI*, vol. 4, pp. 52–64, 2025.
- [89] J. Sun, P. Mao, L. Kong, and J. Wang, "A review of embodied grasping," *Sensors*, vol. 25, p. 852, 2025.
- [90] 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 collaboration 0," in *ICRA*, 2024, pp. 6892–6903.
- [91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
- [92] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *ICCV*, 2023, pp. 11975–11986.
- [93] M. Oquab, T. Dariseti, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: learning robust visual features without supervision," *TMLR*, 2024.
- [94] 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 *ICML*, 2024, pp. 20413–20451.
- [95] W. Chen, S. Belkhale, S. Mirchandani, O. Mees, D. Driess, K. Pertsch, and S. Levine, "Training strategies for efficient embodied reasoning," *arXiv:2505.08243*, 2025.
- [96] S. Dey, J.-N. Zaech, N. Nikolov, L. Van Gool, and D. P. Paudel, "Revla: reverting visual domain limitation of robotic foundation models," in *ICRA*, 2025.
- [97] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang, "Tracevla: visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," in *ICLR*, 2025.
- [98] 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," in *ICRA*, 2025.
- [99] 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 *CVPR*, 2025, pp. 22508–22519.
- [100] 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," in *RSS*, 2025.
- [101] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "Upvla: a unified understanding and prediction model for embodied agent," in *ICML*, 2025.
- [102] 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 *ICLR*, 2025.
- [103] 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:2505.09577*, 2025.
- [104] 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:2505.11214*, 2025.
- [105] T. Van Vo, T. Q. Nguyen, K. M. Nguyen, D. H. M. Nguyen, and M. N. Vu, "Refinevla: reasoning-aware teacher-guided transfer fine-tuning," *arXiv:2505.19080*, 2025.
- [106] Y. Yang, J. Sun, S. Kou, Y. Wang, and Z. Deng, "Lohovla: a unified vision-language-action model for long-horizon embodied tasks," *arXiv:2506.00411*, 2025.
- [107] Y. Wang, X. Li, W. Wang, J. Zhang, Y. Li, Y. Chen, X. Wang, and Z. Zhang, "Unified vision-language-action model," *arXiv:2506.19850*, 2025.
- [108] J. Zhang, Y. Chen, Y. Xu, Z. Huang, Y. Zhou, Y.-J. Yuan, X. Cai, G. Huang, X. Quan, H. Xu, and L. Zhang, "4d-vla: spatiotemporal vision-language-action pretraining with cross-scene calibration," *arXiv:2506.22242*, 2025.
- [109] J. Lin, A. Taherin, A. Akbari, A. Akbari, L. Lu, G. Chen, T. Padir, X. Yang, W. Chen, Y. Li *et al.*, "Vote: vision-language-action optimization with trajectory ensemble voting," *arXiv:2507.05116*, 2025.
- [110] M. A. Patratskiy, A. K. Kovalev, and A. I. Panov, "Spatial traces: Enhancing vla models with spatial-temporal understanding," *arXiv:2508.09032*, 2025.
- [111] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, H. Li, and T. Kong, "Vision-language foundation models as effective robot imitators," in *ICLR*, 2024, pp. 1–19.
- [112] 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 *NeurIPS*, vol. 37, 2024, pp. 40085–40110.
- [113] W. Song, J. Chen, P. Ding, H. Zhao, W. Zhao, Z. Zhong, Z. Ge, J. Ma, and H. Li, "Accelerating vision-language-action model integrated with action chunking via parallel decoding," *arXiv:2503.02310*, 2025.
- [114] 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:2503.20384*, 2025.
- [115] X. Tan, Y. Yang, P. Ye, J. Zheng, B. Bai, X. Wang, J. Hao, and T. Chen, "Think twice, act once: token-aware compression and action reuse for efficient inference in vision-language-action models," *arXiv:2505.21200*, 2025.
- [116] H. Wang, C. Xiong, R. Wang, and X. Chen, "Bitvla: 1-bit vision-language-action models for robotics manipulation," *arXiv:2506.07530*, 2025.
- [117] S. Wang, R. Yu, Z. Yuan, C. Yu, F. Gao, Y. Wang, and D. F. Wong, "Spec-vla: speculative decoding for vision-language-action models with relaxed acceptance," *arXiv:2507.22424*, 2025.
- [118] W. Li, R. Zhang, R. Shao, J. He, and L. Nie, "Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification," *arXiv:2508.21046*, 2025.
- [119] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: deep hierarchical feature learning on point sets in a metric space," in *NeurIPS*, vol. 30, 2017.
- [120] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Language conditioned spatial relation reasoning for 3d object grounding," *NeurIPS*, vol. 35, pp. 20522–20535, 2022.
- [121] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," in *CoRL*, 2025, pp. 3157–3181.
- [122] A. Gu and T. Dao, "Mamba: linear-time sequence modeling with selective state spaces," *arXiv:2312.00752*, 2023.
- [123] J. Nielsen and P. Schneider-Kamp, "Bitnet b1. 58 reloaded: state-of-the-art performance also on smaller networks," in *Delta*, 2024, pp. 301–315.
- [124] 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," in *RSS*, 2025.
- [125] B. Han, J. Kim, and J. Jang, "A dual process vla: efficient robotic manipulation leveraging vlm," *arXiv:2410.15549*, 2024.
- [126] Q. Bu, H. Li, L. Chen, J. Cai, J. Zeng, H. Cui, M. Yao, and Y. Qiao, "Towards synergistic, generalized, and efficient dual-system for robotic manipulation," *arXiv:2410.08001*, 2024.
- [127] Y. Shentu, P. Wu, A. Rajeswaran, and P. Abbeel, "From llms to actions: latent codes as bridges in hierarchical robot control," in *IROS*, 2024, pp. 8539–8546.
- [128] 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:2411.19650*, 2024.
- [129] J. Zhang, Y. Guo, X. Chen, Y.-J. Wang, Y. Hu, C. Shi, and J. Chen, "Hirt: enhancing robotic control with hierarchical robot transformers," in *CoRL*, 2025, pp. 933–946.
- [130] 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:2502.14420*, 2025.
- [131] Z. Zhou, Y. Zhu, J. Wen, C. Shen, and Y. Xu, "Chatvla-2: vision-language-action model with open-world embodied reasoning from pretrained knowledge," *arXiv:2505.21906*, 2025.
- [132] J. Wen, Y. Zhu, M. Zhu, Z. Tang, J. Li, Z. Zhou, X. Liu, C. Shen, Y. Peng, and F. Feng, "Diffusionvla: scaling robot foundation models via unified diffusion and autoregression," in *ICML*, 2025.
- [133] Z. Liu, Y. Gu, S. Zheng, X. Xue, and Y. Fu, "Trivla: a unified triple-system-based unified vision-language-action model for general robot control," *arXiv:2507.01424*, 2025.
- [134] S. Li, L. Gao, J. Wang, C. Che, X. Xiao, J. Cao, Y. Hu, and H. R. Karimi, "Information-theoretic graph fusion with vision-language-action model for policy reasoning and dual robotic control," *arXiv:2508.05342*, 2025.

- [135] 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:2506.10826*, 2025.
- [136] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Vq-vla: improving vision-language-action models via scaling vector-quantized action tokenizers," in *ICCV*, 2025.
- [137] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen *et al.*, "Tinyvla: towards fast, data-efficient vision-language-action models for robotic manipulation," *RA-L*, 2025.
- [138] 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:2505.22159*, 2025.
- [139] F. Lin, R. Nai, Y. Hu, J. You, J. Zhao, and Y. Gao, "Onetwovla: a unified vision-language-action model with adaptive reasoning," *arXiv:2505.11917*, 2025.
- [140] 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:2507.09160*, 2025.
- [141] 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:2507.15493*, 2025.
- [142] X. Chen, H. Wei, P. Zhang, C. Zhang, K. Wang, Y. Guo, R. Yang, Y. Wang, X. Xiao, L. Zhao *et al.*, "Villa-x: enhancing latent action modeling in vision-language-action models," *arXiv:2507.23682*, 2025.
- [143] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, W. Zhang, H. Cui, Z. Zhang, and H. Wang, "Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv:2505.03233*, 2025.
- [144] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023, pp. 4195–4205.
- [145] 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 *CoRL*, 2022, pp. 1678–1690.
- [146] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *NeurIPS*, 2017, pp. 1–10.
- [147] A. M. Noam Shazeer, A. D. Krzysztof Maziarz, G. H. Quoc Le, and J. Dean, "Outrageously large neural networks: the sparsely-gated mixture-of-experts layer," in *ICLR*, 2017.
- [148] Z. Wu, Y. Zhou, X. Xu, Z. Wang, and H. Yan, "Momanipvla: transferring vision-language-action models for general mobile manipulation," in *CVPR*, 2025, pp. 1714–1723.
- [149] 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 *CoRL*, 2024.
- [150] D. Wu, J. Fan, J. Zang, G. Wang, W. Yin, W. Li, and B. Jin, "Reinforced reasoning for embodied planning," *arXiv:2505.22050*, 2025.
- [151] C. Wang, F. Xia, W. Yu, T. Zhang, R. Zhang, C. K. Liu, L. Fei-Fei, J. Tan, and J. Liang, "Chain-of-modality: learning manipulation programs from multimodal human videos with vision-language-models," in *ICRA*, 2025.
- [152] Y. Li, Z. Gong, H. Li, X. Huang, H. Kang, G. Bai, and X. Ma, "Robotic visual instruction," in *CVPR*, 2025, pp. 12155–12165.
- [153] S. Liu, J. Du, S. Xiang, Z. Wang, and D. Luo, "Long-horizon embodied planning with implicit logical inference and hallucination mitigation," *arXiv:2409.15658*, 2025.
- [154] Y. Hu, F. Lin, T. Zhang, L. Yi, and Y. Gao, "Look before you leap: unveiling the power of gpt-4v in robotic vision-language planning," *arXiv:2311.17842*, 2023.
- [155] 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 *CVPR*, 2025, pp. 1724–1734.
- [156] 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 instruction following with hierarchical vision-language-action models," in *ICML*, 2025.
- [157] 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:2505.23450*, 2025.
- [158] 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," in *CoRL*, 2025.
- [159] S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, and H. Li, "Instruct2act: mapping multi-modality instructions to robotic actions with large language model," *arXiv:2305.11176*, 2023.
- [160] W. Mao, W. Zhong, Z. Jiang, D. Fang, Z. Zhang, Z. Lan, H. Li, F. Jia, T. Wang, H. Fan *et al.*, "Robomatrix: a skill-centric hierarchical framework for scalable robot task planning and execution in open-world," *arXiv:2412.00171*, 2024.
- [161] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: injecting the 3d world into vision-language-action models," *arXiv:2503.07511*, 2025.
- [162] Y. Yuan, H. Cui, Y. Chen, Z. Dong, F. Ni, L. Kou, J. Liu, P. Li, Y. Zheng, and J. Hao, "From seeing to doing: bridging reasoning and decision for robotic manipulation," *arXiv:2505.08548*, 2025.
- [163] K. Zhang, R. Xu, P. Ren, J. Lin, H. Wu, L. Lin, and X. Liang, "Robridge: a hierarchical architecture bridging cognition and execution for general robotic manipulation," in *ICCV*, 2025.
- [164] S. Han, B. Qiu, Y. Liao, S. Huang, C. Gao, S. Yan, and S. Liu, "Robocerebra: a large-scale benchmark for long-horizon robotic manipulation evaluation," *arXiv:2506.06677*, 2025.
- [165] 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:2502.20900*, 2025.
- [166] 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 *CoRL*, 2023.
- [167] Z. Liang, Y. Mu, H. Ma, M. Tomizuka, M. Ding, and P. Luo, "Skilldiffuser: interpretable hierarchical planning via skill abstractions in diffusion-based task execution," in *CVPR*, 2024, pp. 16467–16476.
- [168] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao, "RT-affordance: reasoning about robotic manipulation with affordances," in *CoRL*, 2024.
- [169] H. Wu, H. Zhang, P. Zhang, J. Wang, and C. Wang, "Hibernac: hierarchical brain-emulated robotic neural agent collective for disentangling complex manipulation," *arXiv:2506.08296*, 2025.
- [170] D. Niu, Y. Sharma, G. Biamby, J. Quenum, Y. Bai, B. Shi, T. Darrell, and R. Herzig, "Llarva: vision-action instruction tuning enhances robot learning," in *CoRL*, 2024.
- [171] H. Singh, R. J. Das, M. Han, P. Nakov, and I. Laptev, "Malmm: multi-agent large language models for zero-shot robotics manipulation," in *CVPR*, 2024.
- [172] J. Bi, K. Y. Ma, C. Hao, M. Z. Shou, and H. Soh, "Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback," *arXiv:2507.17294*, 2025.
- [173] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning," *arXiv:2501.12948*, 2025.
- [174] B. Zhao, Z. Wang, J. Fang, C. Gao, F. Man, J. Cui, X. Wang, X. Chen, Y. Li, and W. Zhu, "Embodied-r: collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning," *arXiv:2504.12680*, 2025.
- [175] 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 *ICCV*, 2023, pp. 4015–4026.
- [176] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv:2410.21276*, 2024.
- [177] Z. Zhang, K. Zheng, Z. Chen, J. Jang, Y. Li, S. Han, C. Wang, M. Ding, D. Fox, and H. Yao, "Grape: generalizing robot policy via preference alignment," *arXiv:2411.19309*, 2024.
- [178] J. Liu, F. Gao, B. Wei, X. Chen, Q. Liao, Y. Wu, C. Yu, and Y. Wang, "What can rl bring to vla generalization? an empirical study," *arXiv:2505.19789*, 2025.
- [179] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022, pp. 27730–27744.
- [180] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv:2412.16720*, 2024.
- [181] 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:2505.18719*, 2025.

- [182] Z. Chen, R. Niu, H. Kong, and Q. Wang, "Tgrpo: fine-tuning vision-language-action model via trajectory-wise group relative policy optimization," *arXiv:2506.08440*, 2025.
- [183] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv:2303.08774*, 2023.
- [184] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," in *ICLR*, 2022.
- [185] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *ICML*, 2018, pp. 1861–1870.
- [186] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *arXiv:2410.21845*, 2024.
- [187] 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:2502.05450*, 2025.
- [188] C. Xu, Q. Li, J. Luo, and S. Levine, "Rldg: robotic generalist policy distillation via reinforcement learning," in *RSS*, 2025.
- [189] 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," in *ICRA*, 2025.
- [190] Y. Yang, Y. Wang, Z. Wen, L. Zhongwei, C. Zou, Z. Zhang, C. Wen, and L. Zhang, "Efficientvla: training-free acceleration and compression for vision-language-action models," *arXiv:2506.10100*, 2025.
- [191] 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:2502.02175*, 2025.
- [192] Y. Li, Y. Meng, Z. Sun, K. Ji, C. Tang, J. Fan, X. Ma, S. Xia, Z. Wang, and W. Zhu, "Sp-vla: a joint model scheduling and token pruning approach for vla model acceleration," *arXiv:2506.12723*, 2025.
- [193] K. Ji, Y. Meng, H. Cui, Y. Li, S. Hua, L. Chen, and Z. Wang, "Block-wise adaptive caching for accelerating diffusion policy," *arXiv:2506.13456*, 2025.
- [194] 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:2505.06111*, 2025.
- [195] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, L. Liden, K. Lee, J. Gao, L. Zettlemoyer, D. Fox, and M. Seo, "Latent action pretraining from videos," in *ICLR*, 2025.
- [196] H. He, C. Bai, L. Pan, W. Zhang, B. Zhao, and X. Li, "Learning an actionable discrete diffusion policy via large-scale actionless video pre-training," in *NeurIPS*, 2024, pp. 31124–31153.
- [197] 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," in *ICML*, 2024, pp. 61229–61245.
- [198] 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:2502.14795*, 2025.
- [199] S. Patel, S. Mohan, H. Mai, U. Jain, S. Lazebnik, and Y. Li, "Robotic manipulation by imitating generated videos without physical demonstrations," *arXiv:2507.00990*, 2025.
- [200] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: unified 6d pose estimation and tracking of novel objects," in *CVPR*, 2024, pp. 17868–17879.
- [201] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sirha, A. Zholus *et al.*, "V-jepa 2: self-supervised video models enable understanding, prediction and planning," *arXiv:2506.09985*, 2025.
- [202] S. Tan, K. Dou, Y. Zhao, and P. Kraehenbuehl, "Interactive post-training for vision-language-action models," *arXiv:2505.17016*, 2025.
- [203] Z. Wang, S. Cai, Z. Mu, H. Lin, C. Zhang, X. Liu, Q. Li, A. Liu, X. Ma, and Y. Liang, "Omnijarvis: unified vision-language-action tokenization enables open-world instruction following agents," in *NeurIPS*, 2024.
- [204] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliher *et al.*, "π0. 5: a vision-language-action model with open-world generalization," *arXiv:2504.16054*, 2025.
- [205] 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 *CoRL*, 2022, pp. 991–1002.
- [206] 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 *RSS*, 2022, pp. 1–7.
- [207] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: a comprehensive robotic dataset for learning diverse skills in one-shot," in *ICRA*, 2024, pp. 653–660.
- [208] 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 *RSS*, 2024, pp. 1–26.
- [209] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, W. Ai, B. Martinez *et al.*, "Behavior1k: a human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation," in *CoRL*, 2022.
- [210] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: a benchmark for interpreting grounded instructions for everyday tasks," in *CVPR*, 2020, pp. 10740–10749.
- [211] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: the robot learning benchmark & learning environment," *RA-L*, vol. 5, pp. 3019–3026, 2020.
- [212] M. Grotz, M. Shridhar, T. Asfour, and D. Fox, "Peract2: benchmarking and learning for robotic bimanual manipulation tasks," *arXiv:2407.00278*, 2024.
- [213] 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 *CoRL*, 2020, pp. 1094–1100.
- [214] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman, "Relay policy learning: solving long-horizon tasks via imitation and reinforcement learning," in *CoRL*, 2019.
- [215] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: benchmarking knowledge transfer for lifelong robot learning," in *NeurIPS*, vol. 36, 2023, pp. 44776–44791.
- [216] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *RA-L*, vol. 7, pp. 7327–7334, 2022.
- [217] E. Cherepanov, N. Kachaev, A. K. Kovalev, and A. I. Panov, "Memory, benchmark & robots: a benchmark for solving complex tasks with reinforcement learning," *arXiv:2502.10550*, 2025.
- [218] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating real-world robot manipulation policies in simulation," in *CoRL*, 2024.
- [219] 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 *ICCV*, 2019, pp. 9339–9347.
- [220] 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 *NeurIPS*, vol. 34, 2021, pp. 251–266.
- [221] 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," *ICLR*, 2024.
- [222] 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 *CVPR*, 2020, pp. 11097–11107.
- [223] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: a benchmark for evaluating generalization for robotic manipulation," in *RSS*, 2024.
- [224] S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, Q. Gao, Z. Fei, Z. Yin, Z. Wu, Y.-G. Jiang *et al.*, "Vlabench: a large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks," in *ICCV*, 2025.
- [225] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan *et al.*, "Ego4d: around the world in 3,000 hours of egocentric video," in *CVPR*, 2022.
- [226] 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 *CVPR*, 2024, pp. 19383–19400.
- [227] Y. Chen, Y. Ge, Y. Ge, M. Ding, B. Li, R. Wang, R. Xu, Y. Shan, and X. Liu, "Egoplan-bench: benchmarking egocentric embodied planning with multimodal large language models," *arxiv:2312.06722*, 2024.

- [228] X. Wang, K. Zhao, F. Liu, J. Wang, G. Zhao, X. Bao, Z. Zhu, Y. Zhang, and X. Wang, "Egovid-5m: a large-scale video-action dataset for egocentric video generation," *arXiv:2411.08380*, 2024.
- [229] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: the epic-kitchens dataset," in *ECCV*, 2018.
- [230] K. Maeda, T. Hirasawa, A. Hashimoto, J. Harashima, L. Rybicki, Y. Fukasawa, and Y. Ushiku, "Com kitchens: an unedited overhead-view video dataset as a vision-language benchmark," in *ECCV*, 2024, pp. 123–140.
- [231] C. Fan, "Egovqa-an egocentric video question answering benchmark dataset," in *ICCV*, 2019.
- [232] B. Jia, T. Lei, S.-C. Zhu, and S. Huang, "Egotaskqa: understanding human tasks in egocentric videos," in *NeurIPS*, vol. 35, 2022, pp. 3343–3360.
- [233] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "Egodex: learning dexterous manipulation from large-scale egocentric video," *arXiv:2505.11709*, 2025.
- [234] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: scalable and portable mocap data collection system for dexterous manipulation," in *RSS*, 2024.
- [235] 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:2410.24221*, 2024.
- [236] 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:2503.13441*, 2025.
- [237] J. Yu, L. Fu, H. Huang, K. El-Refai, R. A. Ambrus, R. Cheng, M. Z. Irshad, and K. Goldberg, "Real2render2real: scaling robot data without dynamics simulation or robot hardware," *arXiv:2505.09601*, 2025.
- [238] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *CVPR*, 2018, pp. 1–10.
- [239] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: visual question answering in interactive environments," in *CVPR*, 2018, pp. 4089–4098.
- [240] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *CVPR*, 2019, pp. 6309–6318.
- [241] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied question answering in photorealistic environments with point cloud perception," in *CVPR*, 2019, pp. 6659–6668.
- [242] M. M. Islam, A. Gladstone, R. Islam, and T. Iqbal, "Eqqa-mx: embodied question answering using multimodal expression," in *ICLR*, 2023.
- [243] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud *et al.*, "Openeqqa: embodied question answering in the era of foundation models," in *CVPR*, 2024, pp. 16488–16498.
- [244] J.-W. Choi, Y. Yoon, H. Ong, J. Kim, and M. Jang, "Lota-bench: benchmarking language-oriented task planners for embodied agents," in *ICLR*, 2024.