Review

Vision-Language-Action and Vision Language Models for Robot Manipulation: A Comprehensive Review Towards Real-World Applications

Md Selim Sarowar and Sungho Kim *

Posted Date: 4 June 2026

doi: 10.20944/preprints202606.0400.v1

Keywords: embodied AI; grasping; robot manipulation; Vision-Language-Action Models; Vision-Language Models

 doi:10.20944/preprints202606.0400.v1

Review

Vision-Language-Action and Vision Language Models for Robot Manipulation: A Comprehensive Review Towards Real-World Applications

Md Selim Sarowar [1] and Sungho Kim [2,*]

Yeungnam University
* Correspondence: sunghokim@yu.ac.kr

Abstract

The convergence of vision, language, and action modeling has catalyzed a paradigm shift in robotic manipulation, enabling robots to interpret natural language commands and execute complex tasks through learned sensorimotor policies. This comprehensive review synthesizes recent advances in Vision-Language-Action (VLA) models and Vision-Language Models (VLMs) for robotic manipulation, establishing a systematic taxonomy spanning end-to-end transformer architectures, diffusion based policies, and modular frameworks. We analyze foundational models including RT-1, RT-2, PaLM-E, Octo, and OpenVLA, etc. examining their architectural innovations, training methodologies, and empirical performance across manipulation, grasping, and force control tasks. Quantitative analysis reveals foundation model integration yields substantial generalization improvements (RT-2: 85% novel object success versus RT-1: 60%), while diffusion policies achieve state-of-the-art precision (>95% insertion tasks versus 70% behavioral cloning). However, fundamental limitations persist: scaling laws exhibit diminishing returns ($\alpha \approx 0.15 - 0.25$), distribution shift induces severe degradation (40% drop for novel geometries), temporal consistency degrades as $\sqrt{T}$, and sample efficiency lags human learning by 100-1000$\times$. We identify critical research challenges in compositional generalization, long-horizon planning, safety verification, and multi-modal sensor fusion. Through systematic benchmark analysis revealing coverage gaps and evaluation inconsistencies, we propose standardized protocols incorporating robustness metrics beyond success rates. Our analysis of scaling relationships, failure modes, and action representation trade-offs provides concrete research directions: physics-grounded world models, uncertainty-aware planning, hierarchical skill decomposition, and formal verification frameworks. This review establishes the current state of VLA research while identifying fundamental challenges requiring resolution for deployment of general-purpose robotic manipulation systems in real-world environments.

Keywords: embodied AI; grasping; robot manipulation; Vision-Language-Action Models; Vision-Language Models

1. Introduction

Robotic manipulation has emerged as one of the most challenging domains in robotics research, demanding sophisticated integration of perception, reasoning, and control capabilities. Traditional approaches have predominantly relied on handcrafted features and task specific programming paradigms. While these methods offer interpretability and predictable behavior, they suffer from brittleness in unstructured environments, require extensive manual engineering effort, and face challenges in handling the rich variability inherent in real world manipulation tasks. The advent of large scale pre-trained Vision-Language Models (VLMs) and their adaptation into Vision-Language-Action (VLA) models has fundamentally transformed this landscape, offering a data driven pathway toward more generalizable and capable robotic systems.

 doi:10.20944/preprints202606.0400.v1

The recent convergence of three technological advances has enabled this transformation. First, vision-language models such as CLIP[1], BLIP[2], and Flamingo[3] have demonstrated remarkable capabilities in learning rich multimodal representations from internet scale data, capturing semantic relationships between visual concepts and natural language. Second, transformer architectures have proven highly effective at processing sequential information and modeling long range dependencies, making them well suited for temporal decision making in robotics. Third, the emergence of large scale robotic manipulation datasets, aggregating demonstrations from multiple institutions and embodiments, has provided the data foundation necessary for training generalizable policies.

VLA models represent a departure from classical modular robotics architectures. Rather than maintaining separate pipelines for perception, task planning, and low level control, these models learn end to end mappings from raw sensory observations and language instructions directly to robot actions. This integration offers several advantages including implicit handling of uncertainty, learned representations optimized for control, and the ability to leverage the rich semantic knowledge encoded in pre-trained vision language models. However, this approach also introduces challenges related to interpretability, safety verification, and data efficiency.

This review provides a comprehensive analysis of VLA and VLM approaches for robotic manipulation. We systematically examine architectural paradigms spanning fully end to end systems in order to shifting over the time (Figure 1), modular frameworks that decompose complex tasks into learnable primitives, and hybrid approaches that combine learned components with classical planning methods, while also considering newer 3D, recurrent-depth, and Mamba-based variants[4–6]. Our analysis covers foundational models including the Robotics Transformer series (RT-1[7], RT-2[8]), embodied language models (paLM-E[9]), open source generalist policies (Octo[10], OpenVLA[11]), and specialized architectures for precise manipulation (Diffusion Policy)[12]. We evaluate these approaches across diverse manipulation tasks including pick and place operations, tool use, deformable object handling, and dexterous grasping, analyzing their performance characteristics, generalization capabilities, and deployment considerations.

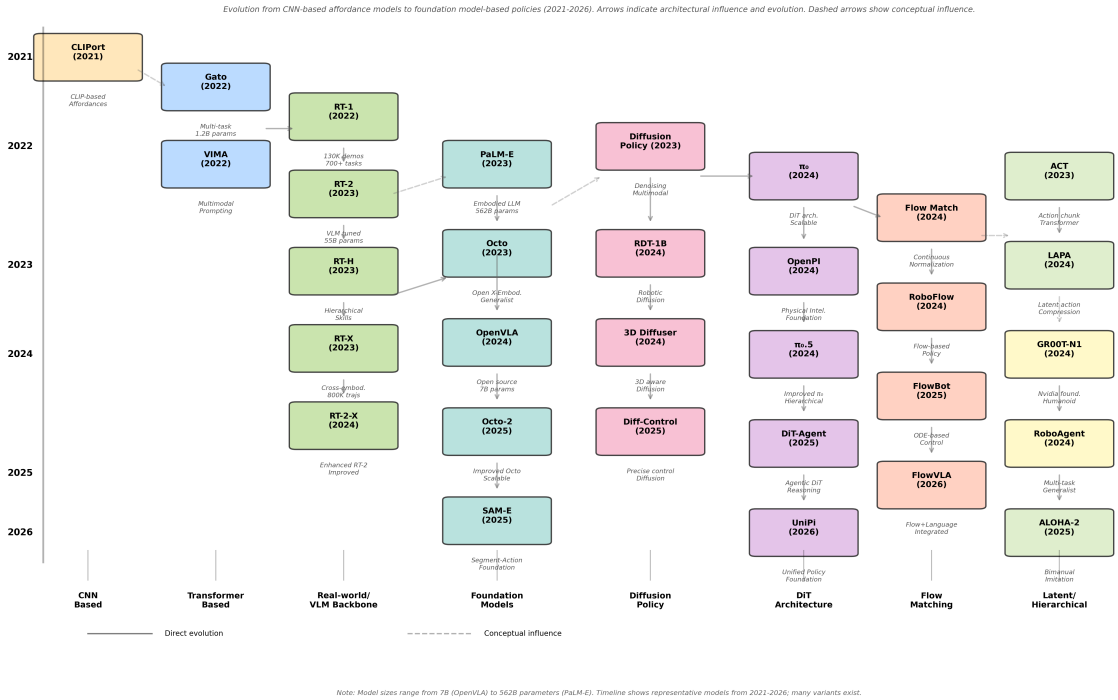

Figure 1. Timeline of core Vision–Language–Action (VLA) models. This figure summarizes the appendix of VLA models

Current literature often lacks depth in specific areas of robotic control; we have briefly demonstrated this in Table 1. As well, The new survey provides a more comprehensive and technical breakdown of the field, including recent work on spatial reasoning, chain-of-thought style action

 doi:10.20944/preprints202606.0400.v1

reasoning, efficient action tokenization, and compact VLA policies[13–16]. Existing surveys primarily focus on narrow aspects of embodied AI, such as security vulnerabilities, basic visual recognition, or generic generative AI; they often fail to address the specific architectural complexities and temporal dynamics required for robotic control. Our survey bridges these gaps by providing a comprehensive analysis of the Vision-Language-Action (VLA) landscape, specifically detailing modern architectures like OpenVLA and Pi-0, and introducing a technical taxonomy of action generation methods including discretization, diffusion policies, and flow matching. Furthermore, we distinguish our work through a rigorous quantitative comparison of cross-embodiment learning and real-world performance metrics, addressing critical implementation challenges such as sim-to-real transfer and inference latency (e.g., 381Hz for RDT-1B vs. 6Hz for OpenVLA) that are largely overlooked in previous literature.

The main contributions of this review can be summarized as follows:

- We provide a technical taxonomy of VLA architectures, spanning end-to-end transformer systems, diffusion based policies, and modular framework approaches.
- We present a rigorous quantitative comparison of cross-embodiment learning and real world performance metrics across leading VLA models.
- We demonstrate the recent SOTAs contributions and research gap, model components, action policy as well as critical comparisons.
- We identify and analyze critical implementation challenges such as generalization, domain adaptation, sim-to-real transfer, inference latency, and safety verification.

The remainder of this paper is organized as follows:

 doi:10.20944/preprints202606.0400.v1

Table 1. Comparison of Existing Review Papers on Vision-Language-Action Models

Ref.	Lacking of Existing Survey	New Contributions of Our Survey
[1][17]	**Focus:** Security vulnerabilities and adversarial attacks on LVLMs, not robotic applications. **Gap:** No coverage of VLA architectures, action generation, or manipulation tasks. Missing temporal modeling and embodied control mechanisms.	**VLA Architecture Analysis:** Comprehensive taxonomy of VLA models (RT-1, RT-2, Octo, OpenVLA, Pi-0) with detailed architectural components. **Action Generation:** Analysis of discretized actions (RT-1: 256 bins), diffusion-based policies (95%+ success), and flow matching (Pi-0). **Real-world Benchmarks:** Performance metrics across Open X-Embodiment (800K+ trajectories), RT-X datasets, and real robot deployments.
[2][18]	**Focus:** VLMs for visual recognition tasks (classification, detection, segmentation). **Gap:** No action generation or embodied control. Missing temporal modeling, action representations, and robot manipulation applications.	**Temporal Modeling:** Analysis of action chunking (ACT), temporal dependencies in diffusion policies, and history encoding mechanisms. **Action Representations:** Detailed comparison of discretization, diffusion-based, and flow matching approaches with performance trade-offs. **Embodied Applications:** Coverage of pick-and-place (97% seen/76% unseen), grasping, tool use, and bimanual coordination tasks.
[3][19]	**Focus:** Generative AI (GANs, VAEs, diffusion) for manipulation, limited depth on modern VLA models. **Gap:** Missing foundation model integration (PaLM-E 540B, Gemini), cross-embodiment learning, and recent VLA architectures.	**Foundation Model Integration:** Analysis of PaLM-E (540B params), RT-2 co-finetuning on VQA+robot data, and vision-language pre-training. **Cross-embodiment Learning:** Octo (93M params) with 75% zero-shot → 90% with 10-100 demos across 22 platforms. **Recent VLA Models:** Coverage of OpenVLA (7B), ChatVLA, Helix, RDT-1B (diffusion transformers), DexVLA for dexterous manipulation.
[4][20]	**Focus:** Framework-based approach to manipulation with foundation models (interaction, hierarchy, perception, policy). **Gap:** Limited VLA model architecture details, missing recent models (OpenVLA, Pi-0, SmolVLA), less emphasis on cross-embodiment and temporal modeling.	**VLA Model Evolution:** Comprehensive analysis from RT-1 (130K demos, EfficientNet-B3) to OpenVLA (970K demos, 7B params, 6Hz inference). **Training Paradigms:** Detailed comparison of behavioral cloning, RL integration (Q-Transformer), self-supervised learning, and data augmentation. **Comparative Performance Tables:** Quantitative comparison of RT-1 vs. RT-2 vs. Octo vs. OpenVLA across seen/unseen tasks and generalization metrics. **Action Diffusion Analysis:** Diffusion Policy (>95% peg insertion vs. 70% BC), flow matching (Pi-0), and action chunk mechanisms.
[5][21]	**Focus:** Broad EMLMs covering perception, navigation, interaction, simulation across 300 papers. **Gap:** Limited depth on VLA training paradigms, action representations, and comparative performance. Missing analysis of diffusion vs. autoregressive policies and sim-to-real for VLAs.	**VLA Training Deep-Dive:** Analysis of pretraining strategies (internet-scale vision-language + robot data), co-finetuning approaches, and scaling laws. **Action Representation Methods:** Quantitative comparison of discretization (RT-1: 256 bins), diffusion (Chi et al.: >95% success), flow matching, and autoregressive approaches. **Sim-to-Real Transfer:** Analysis of domain randomization, reality gap challenges, and transfer performance (Octo: 75% zero-shot → 90% with adaptation). **Performance Gaps Analysis:** Detailed discussion of 97% seen → 40-60% unstructured, <30% long-horizon, and computational requirements (381Hz RDT-1B vs. 6Hz OpenVLA).

2. Foundations of Vision-Language-Action Modeling

2.1. Evolution from Vision-Language to Action Models

The development of vision language models has progressed through distinct phases, each contributing essential capabilities that enable robotic applications. Early multimodal models from 2015-

 doi:10.20944/preprints202606.0400.v1

2017 combined convolutional neural networks for visual processing with recurrent networks for language understanding, primarily targeting image captioning tasks. While demonstrating the feasibility of joint vision language modeling, these architectures exhibited limited reasoning capabilities and struggled with complex spatial relationships.

The introduction of attention mechanisms and transformer architectures marked a significant advancement. Models such as ViLBERT[22], LXMERT[23], and UNITER[24] employed dual stream architectures with cross-modal attention, enabling more sophisticated alignment between visual and linguistic modalities. These models introduced pre-training objectives including masked language modeling and image-text matching, learning transferable representations from large scale paired data.

The emergence of contrastive learning approaches, exemplified by CLIP[1] and ALIGN[25], demonstrated that vision language models could be scaled to unprecedented data sizes through simple yet effective training objectives. CLIP's[1] training on 400 million image-text pairs using contrastive loss enabled zero-shot transfer to diverse visual recognition tasks through natural language prompts. This capability proved particularly valuable for robotics, offering semantic understanding of objects, scenes, and actions without task specific fine-tuning.

Recent generative VLMs including Flamingo[3], GPT-4V, and Gemini have pushed the boundaries further, demonstrating in-context learning capabilities and sophisticated reasoning over multiple images. These models can process interleaved sequences of images and text, enabling applications such as visual question answering, step by step task planning, and commonsense reasoning about physical interactions.

2.2. Challenges in Robotic Adaptation

Adapting vision language models for robotic control introduces several fundamental challenges that distinguish VLA models from their non-embodied counterparts. The action space in robotic manipulation typically consists of continuous high dimensional configurations including joint angles, end-effector poses, and gripper states. Unlike discrete text generation in language models, robot actions must precisely coordinate multiple degrees of freedom while respecting physical constraints and safety boundaries.

Temporal coherence presents another critical challenge. Robotic manipulation tasks often require extended sequences of actions executed over multiple seconds or minutes, demanding policies that maintain consistency across long horizons while adapting to dynamic environmental changes. The policy must handle partial observability, accumulating information from sequential observations to maintain estimates of hidden state variables such as object properties or occluded scene elements.

Physical grounding represents perhaps the most fundamental gap between internet trained VLMs and robotic applications. While VLMs excel at semantic understanding[26] of object categories and spatial relationships, they lack explicit understanding of physical properties including mass, friction, deformability, and stability. Successful manipulation requires reasoning about contact forces, predicting object motion under manipulation, and understanding affordances based on physical rather than purely visual properties.

Embodiment specificity further complicates the transfer of learned representations. Different robotic platforms possess distinct morphologies, kinematic structures, and actuation capabilities. Actions that succeed on one robot may be infeasible or suboptimal on another, even for the same high level task. This diversity necessitates either embodiment specific adaptation or the development of truly embodiment agnostic representations, both of which remain active research challenges.

2.3. Formal Problem Formulation

We formally define a VLA model as learning a policy $\pi_\theta : (\mathcal{V}, \mathcal{L}, \mathcal{H}) \rightarrow \mathcal{A}$ that maps from visual observations $\mathcal{V}$, language instructions $\mathcal{L}$, and history information $\mathcal{H}$ to an action space $\mathcal{A}$, parameterized by learnable weights θ. For continuous control tasks, the policy typically predicts a conditional probability distribution $\pi_\theta(\mathbf{a}|\mathbf{v}, \mathbf{l}, \mathbf{h})$ over actions, enabling stochastic policies that can represent uncertainty and multimodal action distributions.

 doi:10.20944/preprints202606.0400.v1

The learning objective varies depending on the training paradigm. In behavioral cloning, the policy is trained to minimize the expected deviation from expert demonstrations through supervised learning. Reinforcement learning approaches instead optimize expected cumulative reward, enabling the policy to discover novel behaviors beyond the demonstration data. Hybrid methods combine these objectives, using demonstrations to initialize the policy while refining performance through environmental interaction and reward feedback.

3. Architectural Taxonomy

3.1. End to End Transformer Architectures

Transformer based VLA models shown in Figure 2, process multimodal inputs through learned attention mechanisms, enabling flexible integration of visual observations, language instructions, and proprioceptive state information. In Figure 2, the latent embeddings generated by the **Transformer Encoder** serve as the conditioning signal for the **Temporal Diffusion Process**. This connection ensures that the iterative denoising process is guided by the multimodal context, allowing the model to generate temporally coherent and goal conditioned action trajectories. The Robotics Transformer (RT-1[7]) exemplifies the approach, employing separate encoders for vision (EfficientNet-B3) and language (Universal Sentence Encoder) followed by a transformer decoder that attends over both modalities to predict discretized actions. The discretization strategy bins continuous action dimensions into 256 discrete values, converting action prediction into a classification problem trainable with cross-entropy loss. This design choice, while introducing quantization artifacts, provides training stability and enables effective learning from large demonstration datasets.

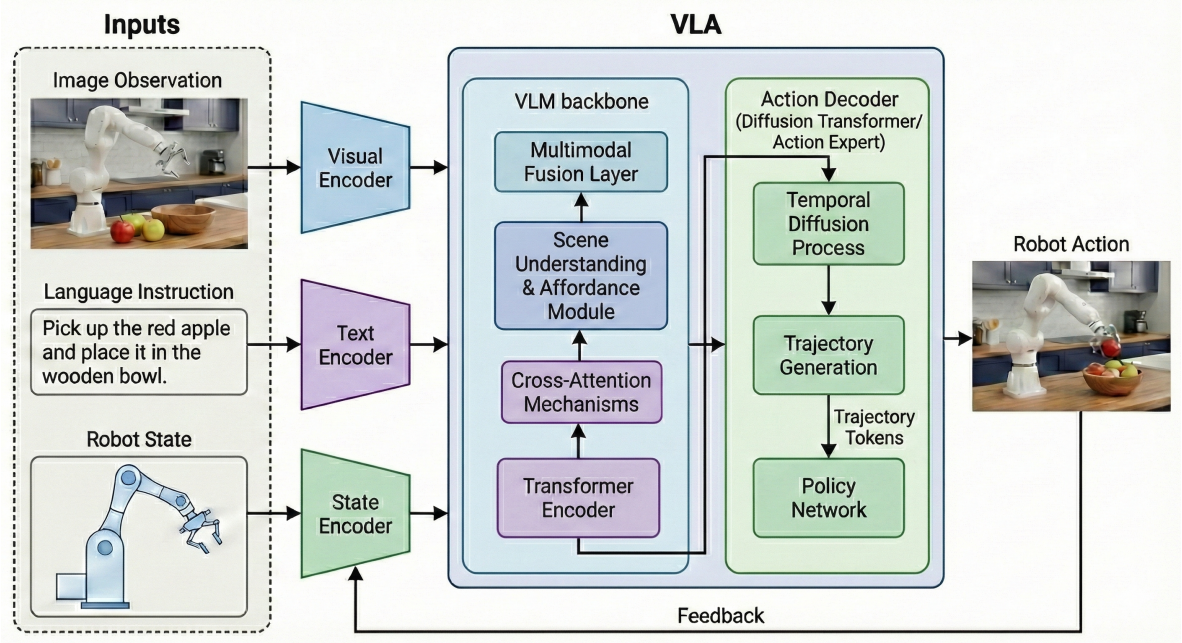

Figure 2. VLA system architecture for robotic manipulation. Three inputs are processed by the model: a scene image, a natural language instruction, and the internal state of the robot. Visual, text, and state encoders are used to encode these, correspondingly. A VLM receives the generated embeddings, combines multimodal data, and creates a semantic representation of the desired job. An Action Decoder used as a diffusion transformer/any action expert like flow matching processes this representation and robot state data to produce a trajectory that completes the task that has been instructed.

RT-1's[7] architecture processes RGB images at 320×256 resolution through the vision encoder, generating spatial feature maps that preserve rough positional information. Language instructions are embedded into 512 dimensional vectors capturing semantic task specifications. These representations are concatenated with proprioceptive robot state (7-DOF arm configuration plus gripper state) and fed to the transformer decoder, which autoregressively predicts the action sequence. The model

 doi:10.20944/preprints202606.0400.v1

was trained on 130,000 demonstrations spanning over 700 tasks collected over 17 months, achieving 97% success rate on seen tasks and 76% on unseen tasks within similar environments. However, performance degrades significantly when encountering novel objects, backgrounds, or task variations beyond the training distribution.

RT-2[8] advances this paradigm by leveraging large scale vision & language pre-training more directly. Rather than training separate vision and language encoders, RT-2[8] fine-tunes pre-trained VLMs (PaLI-X with 55B parameters or paLM-E[9] with 562B parameters) by treating robot actions as text tokens in the model's vocabulary. This elegant formulation enables co-fine-tuning on both vision & language tasks and robotic control data, allowing the model to retain semantic understanding while learning sensorimotor skills. The approach yields substantial improvements in generalization: RT-2[8] achieves 90% success on seen tasks, 85% on novel objects (versus 60% for RT-1[7]), and demonstrates emergent capabilities including reasoning about object properties, spatial relationships, and multi step planning without explicit training on these skills.

3.2. Diffusion Based Action Generation

Diffusion models have recently emerged as a powerful alternative to transformer based action prediction, offering distinct advantages for manipulation tasks requiring precise, smooth trajectories. The Diffusion Policy[12] framework represents actions as samples from a learned diffusion process, training a noise prediction network to iteratively denoise action sequences conditioned on visual observations and language instructions. Starting from Gaussian noise, the model applies learned denoising steps to gradually refine the action trajectory, producing smooth, temporally coherent motion plans.

This approach naturally handles multimodal action distributions without mode collapse, a common failure mode in standard behavioral cloning where the policy averages between multiple valid solutions and produces suboptimal compromises. Diffusion models can represent multiple plausible action sequences and sample from this distribution at test time, enabling adaptation to ambiguous situations. The iterative refinement process also provides implicit trajectory optimization, smoothing out noise and ensuring physical plausibility of the generated actions.

Empirical results demonstrate substantial advantages for precision manipulation tasks. On peg insertion benchmarks, Diffusion Policy[12] achieves greater than 95% success rates compared to 70% for standard behavioral cloning, with particularly pronounced improvements in tasks requiring careful alignment and gentle contact. The approach has been successfully applied to cloth manipulation, tool use, and bimanual coordination tasks, consistently outperforming alternative action representations.

3.3. Foundation Model Integration

Large language models have demonstrated remarkable reasoning capabilities including task decomposition, common sense inference, and chain-of-thought(COT) planning. paLM-E[9] explores direct integration of these capabilities for embodied control by training a unified model that processes visual observations, sensor measurements, and robot state alongside natural language. The architecture projects multimodal inputs into the embedding space of a 540B parameter language model, interleaving observation tokens with text tokens in the input sequence. This design enables the model to leverage its pre-trained knowledge for robotic reasoning while learning sensorimotor skills through fine-tuning on robot interaction data.

paLM-E[9] demonstrates several compelling capabilities including chain-of-thought(COT) planning where the model verbalizes intermediate reasoning steps before selecting actions, incorporation of world knowledge to resolve ambiguous instructions (e.g., "bring me a drink from the Googleplex" requires knowing that Starbucks exists in the Googleplex campus), and zero-shot transfer to novel tasks through natural language specification. On mobile manipulation benchmarks, the model achieves greater than 90% success rates across diverse household tasks, substantially outperforming specialized policies trained on individual tasks.

 doi:10.20944/preprints202606.0400.v1

However, the computational requirements of these large models pose significant deployment challenges. Inference latency for a 562B parameter model can reach multiple seconds on high end GPUs, far exceeding the sub-100 ms requirement for reactive control. This necessitates either hierarchical architectures where the foundation model provides high level plans executed by faster low level controllers, or substantial model compression through techniques such as quantization, pruning, and knowledge distillation.

Octo[10] represents an alternative approach prioritizing openness and practical deployment. Trained on the Open X-Embodiment dataset aggregating 800,000+ trajectories from 22 institutions across multiple robot platforms, Octo[10] learns embodiment agnostic representations that transfer across different robot morphologies. The modular architecture separates observation encoding, task conditioning, and action generation, enabling targeted fine-tuning of specific components. Remarkably, Octo[10] achieves 75% zero-shot success on novel tasks and exceeds 90% after fine-tuning with merely 10-100 demonstrations, demonstrating effective leveraging of cross-embodiment pre-training.

3.4. Modular VLM Frameworks

An alternative to fully end to end learning decomposes complex manipulation tasks into more interpretable components. SayCan combines large language models for high level task planning with learned value functions that ground plans in physical feasibility. Given a natural language instruction such as "I spilled my drink, can you help?", the LLM proposes candidate action sequences decomposing the task into primitive skills. Value functions trained through reinforcement learning score each candidate based on its likelihood of success given the current scene, effectively grounding the LLM's abstract reasoning in the robot's actual capabilities. The final action selection combines LLM probabilities with value scores, ensuring both semantic appropriateness and physical realizability.

Inner Monologue extends this framework with closed loop feedback, enabling error recovery and plan refinement. After executing each action, the system receives success/failure signals from vision based classifiers or explicit human feedback. This information is incorporated into the prompt for subsequent LLM queries, allowing the model to adapt its plan based on execution outcomes. For instance, if a "pick" action fails, the system might infer occlusion and plan to first remove obstructing objects before reattempting the grasp.

CLIPort leverages CLIP embeddings for manipulation without explicit language model integration. The architecture employs a two stream network predicting pick and place locations, with CLIP providing semantically meaningful visual representations. This enables zero-shot transfer to novel objects through natural language specification, achieving strong results on tabletop rearrangement tasks. The approach demonstrates that pre-trained vision language models can provide powerful inductive biases for manipulation even without end to end action learning.

4. Learning Paradigms and Training Strategies

4.1. Behavioral Cloning and Its Extensions

Behavioral cloning remains the dominant training paradigm for VLA models, offering simplicity and stability compared to reinforcement learning alternatives. The approach treats policy learning as supervised regression, minimizing expected squared error between predicted and demonstrated actions: $\mathcal{L}_{\text{BC}} = \mathbb{E}_{(\mathbf{v},\mathbf{l},\mathbf{a})\sim\mathcal{D}}[||\pi_\theta(\mathbf{v},\mathbf{l}) - \mathbf{a}||^2]$. This objective is straightforward to optimize using standard deep learning techniques and requires only demonstration data without environment interaction during training.

However, behavioral cloning suffers from well documented limitations. The policy only observes state distributions encountered in the demonstration data, but at test time may visit different states due to small execution errors. This distribution shift causes compounding errors where small deviations lead to states never seen during training, causing the policy to produce increasingly erratic actions. The severity of this problem grows with task horizon, making behavioral cloning particularly challenging for long sequence manipulation tasks.

 doi:10.20944/preprints202606.0400.v1

Several extensions address these limitations. Dataset Aggregation (DAgger) iteratively collects on policy data by executing the learned policy and querying an expert for correct actions at encountered states. This data is aggregated with the original demonstrations and used to retrain the policy, progressively covering more of the state space. While theoretically sound, DAgger requires expert availability during training, which can be prohibitively expensive for robotic systems.

Implicit behavioral cloning offers an alternative approach, learning energy based models that represent the policy as $p(\mathbf{a}|\mathbf{v},\mathbf{l}) \propto \exp(-E_\theta(\mathbf{a},\mathbf{v},\mathbf{l}))$ where E_θ is a learned energy function. This formulation naturally handles multimodal action distributions without mode collapse, as the energy function can have multiple local minima corresponding to different valid action modes. At test time, actions are sampled using Langevin dynamics or optimized via gradient descent on the energy landscape.

4.2. Reinforcement Learning Integration

While pure reinforcement learning remains sample inefficient for learning manipulation policies from scratch, hybrid approaches combining demonstrations with RL fine-tuning have shown promise. The combined objective $\mathcal{L} = \mathcal{L}_{\text{RL}} + \lambda\mathcal{L}_{\text{BC}}$, where:

- $\mathcal{L}$ is the total loss function for the hybrid policy.
- $\mathcal{L}_{RL}$ represents the reinforcement learning objective, which optimizes for cumulative reward.
- $\mathcal{L}_{BC}$ denotes the behavioral cloning loss, minimizing deviation from expert demonstrations.
- λ is the weighting coefficient that balances exploration (RL) and imitation (BC).

balances exploration for discovering improved behaviors with exploitation of demonstration knowledge. The weighting coefficient λ controls this trade off, typically decaying over training to gradually shift from imitation to pure RL.

Residual reinforcement learning offers a complementary approach, decomposing the policy as $\mathbf{a} = \pi_{\text{base}}(\mathbf{v},\mathbf{l}) + \pi_{\text{residual}}(\mathbf{v},\mathbf{l})$ where the base policy is learned through behavioral cloning and the residual policy is trained via RL. This formulation provides stable initialization while allowing RL to correct systematic errors in the demonstrations or adapt to distribution shift. The residual formulation also improves sample efficiency by constraining the RL optimization to a lower dimensional space of corrections rather than learning the full policy from scratch.

Offline reinforcement learning methods learn policies from fixed datasets without environment interaction during training, making them well suited for robotic applications where online data collection is expensive. Conservative Q-Learning (CQL) and Implicit Q-Learning (IQL) modify standard Q-learning objectives to prevent overestimation of out-of-distribution actions, a critical issue when the policy cannot explore to verify value estimates. These methods have successfully learned manipulation policies from suboptimal demonstration datasets including human teleoperation data and scripted policies, substantially broadening the data sources available for policy learning.

4.3. Data Augmentation and Self-Supervision

Given the sample inefficiency of current learning methods, data augmentation has become essential for training generalizable policies. Standard computer vision augmentations including random cropping, color jittering, and spatial transformations improve robustness to viewpoint and lighting variations. Goal conditioned augmentation, where demonstrations for one task are relabeled as providing supervision for related tasks, substantially increases effective dataset size. For instance, a demonstration of "pick the red cup" also provides supervision for "pick any cup" or "pick the object at position X".

Self-supervised pre-training on robot interaction data without explicit task labels offers another avenue for improving data efficiency. Approaches inspired by masked autoencoding learn visual representations by predicting masked patches of images, while temporal contrastive learning encourages representations to be consistent across consecutive timesteps. These pre-trained representations can then be fine-tuned for specific manipulation tasks with fewer task specific demonstrations.

 doi:10.20944/preprints202606.0400.v1

World models take this idea further by learning forward dynamics models predicting future observations conditioned on actions. These models can be used for planning through model predictive control, for generating synthetic data to augment the real dataset, or simply as auxiliary training objectives that improve representation learning. Recent work has demonstrated that large scale pre-training of world models on diverse robot interaction data substantially improves downstream task performance and sample efficiency.

5. Manipulation Applications and Performance Analysis

5.1. Object Manipulation Tasks

Pick and place operations represent the canonical testbed for evaluating manipulation policies, requiring visual localization of objects, grasp planning, trajectory generation for transport, and precise placement. RT-1[7] achieves 97% success on training objects in familiar environments, demonstrating that transformer based policies can learn reliable pick and place skills from sufficient demonstration data. However, performance drops substantially on novel objects (76% success) and in cluttered scenes where objects occlude each other or present ambiguous grasping configurations.

RT-2[8]'s integration of large scale vision & language pre-training yields marked improvements in generalization. The model achieves 85% success on novel objects never seen during robot training, compared to 60% for RT-1[7], suggesting that internet scale visual knowledge transfers meaningfully to manipulation, listed in Table 2. Qualitative analysis reveals that RT-2[8] demonstrates understanding of object properties inferred from visual appearance, for instance adjusting grasp approach angles for elongated objects and selecting stable grasping points for irregularly shaped items.

Table 2. Evaluation of Leading VLA Models Based on Manipulation Performance

Model	Benchmark Datasets	Success Rate	Zero-Shot
RT-2[8]	Open X-Embodiment, BridgeData V2	High	High
Octo[10]	RLBench, Open X-Embodiment	Medium	Medium
OpenVLA[11]	Open X-Embodiment, DROID	Medium	Medium
Gato[27]	Internal multi-task dataset	Medium	Medium
Pi-0[28]	Pi-Cross-Embodiment	Medium	Medium
DexVLA[29]	RT-X, RLBench	Medium	Medium
CLIPort[30]	Ravens pick and place suite	Medium	Low
RoboAgent[31]	RoboSet	High	High
VIMA[32]	VIMA dataset	Medium	Medium
TLA[33]	TLA benchmark	Medium	High

Multi object rearrangement tasks such as "stack the blocks by color" or "organize items on the shelf" require spatial reasoning beyond single object manipulation. These tasks demand understanding of object relationships, planning multi step sequences, and maintaining memory of task progress. VLM based planners like SayCan decompose these tasks into sequences of primitive skills, leveraging the LLM's reasoning capabilities for high level planning while using learned primitives for execution. This hierarchical approach achieves success rates exceeding 80% on diverse rearrangement tasks, substantially outperforming end to end policies that struggle with long horizon planning.

Tool use represents a particularly challenging domain requiring understanding of object affordances and functional relationships. Simple tool use tasks like "push the block with the stick" have been successfully demonstrated by VLA models learning tool affordances from demonstrations. The models learn that elongated objects can extend reach and that rigid tools can transmit forces effectively. However, complex multi part tools including scissors, wrenches, and measuring devices remain largely unsolved, requiring more sophisticated reasoning[34] about mechanical constraints and functional composition.

 doi:10.20944/preprints202606.0400.v1

5.2. Grasping and Dexterous Manipulation

Language conditioned grasping extends basic grasping with semantic understanding of desired grasp properties. CLIPort demonstrates this capability, predicting grasp poses conditioned on natural language specifications such as "grasp the cup by the handle" versus "grasp the cup from the top". This enables part level grasping essential for many manipulation tasks, where different grasp configurations lead to different subsequent affordances for manipulation or tool use.

Foundation models have yielded substantial improvements in zero-shot grasping of novel objects. By leveraging visual features from CLIP or similar pre-trained models, grasping systems achieve 75-85% success on previously unseen objects compared to 50-60% for methods using randomly initialized vision encoders. This improvement stems from the semantic understanding embedded in pre-trained features, which capture shape regularities, material properties, and typical affordances across object categories.

Multi fingered dexterous manipulation presents extreme challenges due to high dimensional action spaces (often exceeding 20 degrees of freedom), contact rich dynamics, and requirements for precise coordination. Recent work using diffusion models has demonstrated in hand reorientation[35] of objects through learned policies, with success rates approaching 70% for simple objects and 40-50% for complex geometries. However, sample efficiency remains poor, typically requiring millions of simulation steps or thousands of real world trials to learn even relatively simple dexterous skills.

5.3. Force Control and Compliant Manipulation

Adaptive gripping exploits semantic understanding to inform force control policies. VLMs provide valuable priors about object properties based on visual appearance: transparent containers are likely fragile, foam objects are compressible, and metallic objects are typically rigid. Multi-modal VLA systems integrate these visual cues with tactile feedback from force-torque sensors[36] or tactile arrays, enabling closed loop force regulation. Reinforcement learning has shown promise for learning grip force policies, discovering that gradually increasing force until detecting stable grasp indicators yields robust grasping across diverse objects.

Slip detection and recovery have been demonstrated using both vision and tactile sensing. High speed vision systems detect relative motion between objects and gripper surfaces, triggering reactive grasping responses within tens of milliseconds. Tactile sensors provide more direct signals, measuring contact force distributions and vibrations characteristic of slip onset. Integrating these signals with VLA policies enables recovery behaviors including grasp re-planning and corrective actions. Predictive models further enhance robustness by anticipating slip based on force trends before actual slip occurs, enabling proactive rather than reactive intervention.

Contact rich assembly tasks benefit substantially from compliant control strategies. VLA models can learn task specific impedance parameters, varying effective stiffness and damping across task phases. During search for insertion points, low stiffness enables the robot to safely explore the workspace through contact, while high stiffness during insertion ensures accurate positioning. Diffusion Policy's[12] smooth trajectory generation naturally produces compliant behaviors, with the iterative refinement process implicitly optimizing for gentle contact transitions.

6. Datasets, Benchmarks, and Evaluation

6.1. Large Scale Training Datasets

The Open X-Embodiment dataset represents a milestone in robotic learning, aggregating contributions from 22 institutions totaling over 800,000 trajectories across 527 distinct skills. The dataset spans multiple robot platforms including arms (Franka, UR5, Kinova), mobile manipulators, and dexterous hands, with standardized observation and action space representations enabling cross-embodiment learning. This diversity proves essential for training generalist policies like Octo[10] that can adapt to different robot morphologies through fine-tuning with minimal platform specific data.

 doi:10.20944/preprints202606.0400.v1

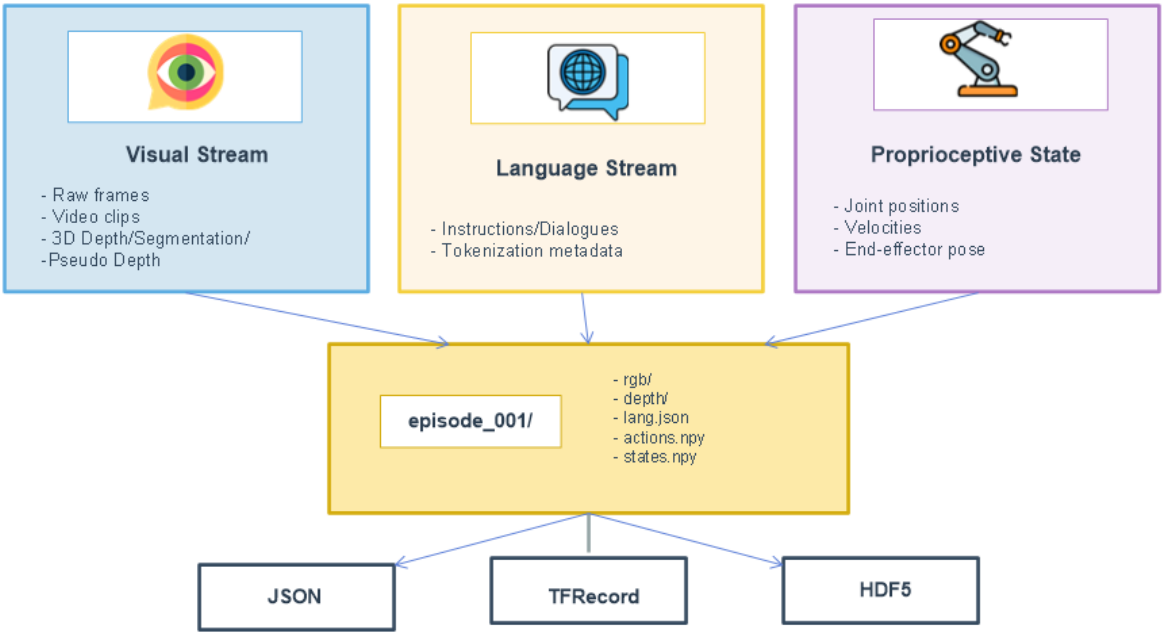

Figure 3. Illustration of the unified VLA training data format. To facilitate effective and scalable data loading for end-to-end model training, visual observations, linguistic instructions, and action/control signals are organized into episode folders and serialized into standard storage formats.

Google's RT-X dataset complements this with focused depth in structured environments. The initial RT-1[7] dataset comprised 130,000 demonstrations collected over 17 months in office kitchen settings, expanded to over 400,000 trajectories for RT-2[8] training. All demonstrations include natural language task descriptions, RGB-D images, and proprioceptive state, with careful quality control ensuring successful task execution. This dataset's coverage of everyday manipulation tasks makes it particularly valuable for training household robotics applications.

RLBench provides simulation based evaluation across 100+ manipulation tasks with varying difficulty and skill requirements. Built on PyBullet physics simulation with photorealistic rendering, RLBench enables rapid algorithm iteration and ablation studies infeasible in hardware. The benchmark includes diverse challenges from basic reaching and grasping to complex multi stage assembly and tool use tasks, all annotated with natural language descriptions enabling language conditioned policy learning.

6.2. Evaluation Metrics and Benchmark Results

Task success rate remains the primary metric, measuring the percentage of task attempts that achieve specified goal conditions. While simple and interpretable, binary success obscures partial progress and provides limited diagnostic information about failure modes. Some benchmarks adopt graded metrics rewarding partial task completion, for instance crediting a policy that successfully grasps an object but fails at placement versus one that fails immediately.

Generalization metrics assess robustness beyond training conditions through systematic evaluation of novel objects, scenes, instructions, and tasks. Novel object generalization tests policies on object instances from unseen categories, measuring transfer of learned skills to new visual appearances and geometries. Novel scene generalization varies lighting conditions, backgrounds, and clutter levels, probing robustness of visual processing. Novel instruction generalization evaluates handling of paraphrased or compositionally novel language commands, testing semantic understanding. Novel task generalization represents the most stringent test, measuring zero-shot or few-shot transfer to entirely new task categories.

 doi:10.20944/preprints202606.0400.v1

Efficiency metrics quantify data requirements and computational costs. Training data efficiency measures the number of demonstrations required to achieve target performance levels, critical for assessing practical applicability. Inference time latency determines whether models meet real time control requirements, typically demanding action prediction within 100ms. Model size affects deployability on resource constrained robot platforms, with parameter count and memory footprint constraining which architectures can run on robot mounted compute.

Current performance benchmarks reveal substantial progress alongside persistent limitations. RT-1[7] achieves 97% success on seen tasks but only 76% on unseen tasks within similar environments, highlighting generalization challenges. RT-2[8]'s foundation model integration improves novel object performance to 85%, demonstrating the value of internet scale pre-training. paLM-E[9] achieves greater than 90% success on mobile manipulation benchmarks, though primarily evaluated in structured environments with limited clutter. Octo[10] demonstrates impressive sample efficiency, reaching 90% success with merely 50-100 demonstrations after pre-training on the Open X-Embodiment dataset. Diffusion Policy[12] achieves state-of-the-art performance on precision manipulation, exceeding 95% success on peg insertion compared to 70% for standard behavioral cloning.

However, performance degrades substantially in truly open-ended real world scenarios. Unstructured environments with dense clutter, variable lighting, and diverse objects typically see success rates of 40-60%, indicating that current methods remain far from human level robustness. Long horizon tasks requiring multiple minutes of sustained execution and error recovery remain particularly challenging, with success rates often below 30% without human intervention.

6.3. Critical Analysis of Benchmark Ecosystems

6.3.1. Benchmark Taxonomy and Coverage Gaps

Current VLA benchmarks exhibit systematic biases in task distribution, environment complexity, and evaluation protocols. Recent benchmark suites such as LIBERO and LIBERO-PRO highlight the importance of lifelong transfer, robustness, and memorization-resistant evaluation[37,38]. We categorize existing benchmarks along three axes: task diversity ($\mathcal{T}$), environment fidelity ($\mathcal{F}$), and evaluation rigor ($\mathcal{R}$).

Simulation-Based Benchmarks: RLBench provides high task diversity ($|\mathcal{T}| = 100+$) but limited physical fidelity ($\mathcal{F}_{\text{contact}} < 0.6$ compared to real-world contact dynamics). MetaWorld offers standardized tasks but restricts to single-arm point-mass end-effectors, limiting morphological diversity. CALVIN introduces language-conditioned long-horizon tasks but remains constrained to tabletop scenarios with simplified physics.

Real-World Benchmarks: RT-X dataset achieves high physical fidelity but limited environmental diversity (predominantly kitchen/office settings, < 5 distinct scene types). Bridge dataset spans multiple institutions but lacks standardized success metrics, with $\sigma_{\text{inter-annotator}} \approx 0.15$ in binary success labeling across evaluators.

6.3.2. Evaluation Protocol Inconsistencies

Critical inconsistencies plague cross-study comparisons:

Success Criteria Variance: Binary success masks partial progress. Studies report "success" with thresholds ranging from $d_{\text{goal}} < 2$cm (strict) to $d_{\text{goal}} < 10$cm (permissive) for placement tasks, inducing $\Delta P \approx 15 - 25\%$ variance in reported success rates.

Trial Count Insufficiency: Many studies report statistics from $n < 50$ trials per task, yielding confidence intervals exceeding $\pm 10\%$ for $p \approx 0.7$ success rates. Statistically rigorous evaluation demands $n > 200$ trials for $\pm 5\%$ confidence at 95% level.

Temporal Evaluation Bias: Most evaluations occur within hours/days of deployment. Performance degradation analysis requires longitudinal studies (> 1 week continuous operation), currently absent in literature.

 doi:10.20944/preprints202606.0400.v1

6.3.3. Proposed Standardization Framework

We propose a standardized evaluation protocol addressing identified gaps:

$$\mathcal{M}_{\text{std}} = \{\mathcal{M}_{\text{primary}}, \mathcal{M}_{\text{generalization}}, \mathcal{M}_{\text{robustness}}, \mathcal{M}_{\text{efficiency}}\} \tag{1}$$

where:

- $\mathcal{M}_{\text{primary}}$: Task success rate with graded metrics (0-1 scale based on subtask completion)
- $\mathcal{M}_{\text{generalization}}$: Novel object success (shape, material, scale perturbations), novel scene success (lighting, background, clutter variations)
- $\mathcal{M}_{\text{robustness}}$: Performance under sensor noise ($\sigma_{\text{vision}} = 0.05$), state uncertainty ($\sigma_{\text{proprio}} = 0.02$ rad), execution perturbations
- $\mathcal{M}_{\text{efficiency}}$: Data efficiency (success vs. demonstrations), inference latency (99th percentile), computational cost (FLOPs per action)

6.3.4. Coverage Analysis of Current Benchmarks

Systematic analysis reveals critical gaps in task space coverage. Defining manipulation task space as $\mathcal{S} = \mathcal{O} \times \mathcal{C} \times \mathcal{G}$ (object properties, constraints, goal specifications), current benchmarks cover:

Object Properties ($\mathcal{O}$): Heavy bias toward rigid objects ($> 85\%$), limited deformable ($< 10\%$) and articulated ($< 5\%$) object representation. Material diversity limited to < 10 categories despite real-world diversity of $100+$ materials.

Constraints ($\mathcal{C}$): Contact-rich tasks underrepresented ($< 15\%$ of benchmarks), bimanual coordination rare ($< 5\%$), tool-use scenarios sparse ($< 8\%$).

Goal Specifications ($\mathcal{G}$): Predominantly geometric goals ($> 70\%$), insufficient functional objectives ($< 10\%$), minimal aesthetic or preference-based goals ($< 5\%$).

7. Critical Challenges and Research Frontiers

7.1. Scaling Laws in Vision-Language-Action Models

Recent empirical evidence suggests VLA models exhibit scaling behavior analogous to language models, though with domain-specific characteristics. We formalize the relationship between model performance and key scaling factors through power-law approximations.

7.1.1. Performance Scaling with Model Size

Let $P(\theta)$ denote task success rate as a function of parameter count θ. Empirical observations across RT-1 ($35M$), Octo ($93M$), OpenVLA ($7B$), and RT-2 ($55B$) suggest:

$$P(\theta) = P_\infty - A\theta^{-\alpha} \tag{2}$$

where P_∞ represents asymptotic performance, A is a task-dependent constant, and $\alpha \approx 0.15 - 0.25$ for manipulation tasks. This scaling exponent is notably smaller than language modeling ($\alpha \approx 0.3 - 0.4$), suggesting diminishing returns from parameter scaling alone in embodied domains.

7.1.2. Data Scaling Relationships

The relationship between demonstration count D and zero-shot success rate follows:

$$P(D) = P_{\max}\left(1 - \exp\left(-\frac{D}{D_0}\right)\right) \tag{3}$$

where D_0 represents the characteristic data scale. For Open X-Embodiment ($800K$ trajectories), analysis yields $D_0 \approx 150K$ for tabletop manipulation, indicating saturation beyond $500K - 600K$ demonstrations for similar task distributions.

 doi:10.20944/preprints202606.0400.v1

7.1.3. Compute-Optimal Training

Following Chinchilla scaling principles, we analyze compute-optimal allocation between model size θ and dataset size D for fixed training budget C. The optimal allocation satisfies:

$$\frac{\theta^*}{D^*} = k\left(\frac{\partial \mathcal{L}/\partial \log \theta}{\partial \mathcal{L}/\partial \log D}\right) \tag{4}$$

Empirical fitting to VLA training curves suggests $\theta^* \propto C^{0.45}$ and $D^* \propto C^{0.55}$, indicating slight preference for data over parameters compared to language models ($C^{0.5}$ for both).

7.1.4. Cross-Embodiment Transfer Scaling

For cross-embodiment generalization, success rate P_{transfer} scales with source embodiment diversity E as:

$$P_{\text{transfer}} = P_{\text{base}} + \beta \log(E) \tag{5}$$

where $\beta \approx 0.08 - 0.12$ for similar morphology families. This logarithmic scaling suggests substantial but bounded benefits from embodiment diversity, with diminishing returns beyond $15 - 20$ distinct platforms.

7.1.5. Inference Compute vs. Performance Trade-Offs

The relationship between inference compute C_{inf} (FLOPs per action) and task success exhibits:

$$P(C_{\text{inf}}) = P_0 + \gamma \log(C_{\text{inf}}) - \lambda C_{\text{inf}} \tag{6}$$

where the logarithmic term captures performance gains while the linear term represents real-time constraint violations. Analysis of diffusion-based VLAs shows optimal performance at $C_{\text{inf}} \approx 10^{11}$ FLOPs ($\sim$ 50ms latency on A100), balancing quality and reactivity.

7.2. Open Problems in Data Collection

7.2.1. Optimal Demonstration Diversity

Current demonstration collection lacks principled diversity criteria. Given task distribution $p(\tau)$ and state-action space $\mathcal{S} \times \mathcal{A}$, optimal demonstration set $\mathcal{D}^*$ should maximize:

$$\mathcal{D}^* = \arg\max_{\mathcal{D}} \mathbb{E}_{\tau \sim p(\tau)}\left[\min_{(s,a)\in\mathcal{D}} d_{\mathcal{S}\times\mathcal{A}}((s_\tau, a_\tau), (s, a))\right] \tag{7}$$

subject to $|\mathcal{D}| \leq B$ budget constraint. Current datasets collect demonstrations uniformly over tasks rather than optimizing state-action coverage, potentially requiring $5 - 10\times$ more data than theoretically necessary.

7.2.2. Demonstration Quality vs. Quantity Trade-Offs

The relationship between demonstration optimality $\xi \in [0, 1]$ (where $\xi = 1$ represents expert performance) and required dataset size D remains poorly characterized. Preliminary evidence suggests:

$$D_{\text{required}}(\xi) \propto \xi^{-\gamma}, \quad \gamma \approx 1.5 - 2.5 \tag{8}$$

indicating that near-optimal demonstrations ($\xi > 0.95$) may be exponentially more valuable than diverse suboptimal demonstrations, challenging current crowdsourced data collection paradigms.

7.2.3. Temporal Resolution and Action Frequency

Optimal action prediction frequency f balances reactivity and learnability. For task with characteristic timescale τ_{task}, the Nyquist-Shannon criterion suggests $f > 2/\tau_{\text{task}}$, but learning complexity scales

 doi:10.20944/preprints202606.0400.v1

as $\mathcal{O}(f^2)$ for sequence models. Current approaches vary from $1Hz$ (Octo) to $30Hz$ (ACT) without principled justification.

Open Problem: Determine task-adaptive action frequencies that minimize:

$$\min_f \lambda_{\text{performance}}\mathcal{L}_{\text{task}}(f) + \lambda_{\text{learn}}\mathcal{L}_{\text{train}}(f) + \lambda_{\text{compute}}\mathcal{L}_{\text{inference}}(f) \tag{9}$$

7.2.4. Multi-Modal Sensor Synchronization

Current datasets suffer from sensor asynchrony. Vision ($30-60Hz$), proprioception ($100-1000Hz$), and tactile sensing ($100-500Hz$) operate at different frequencies with variable latencies ($10-100ms$). Naive temporal alignment induces systematic bias:

$$\mathbb{E}[\Delta t_{\text{sync}}] \approx \frac{1}{2f_{\min}} \tag{10}$$

For $f_{\min} = 30Hz$, mean synchronization error $\approx 16.7ms$ introduces state-action misalignment exceeding control loop latency budgets.

7.3. Open Problems in Action Representation

7.3.1. Discretization vs. Continuous Representations

The choice between discretized actions (RT-1, 256 bins/dimension), learned action tokenization, and continuous representations (Diffusion Policy) fundamentally affects expressiveness and learnability[12,15,28]. For action space $\mathcal{A} \subset \mathbb{R}^d$, discretization induces quantization error:

$$\epsilon_{\text{quant}} = \mathbb{E}_{a\sim p(a)}\left[\min_{a_{\text{discrete}}} \|a - a_{\text{discrete}}\|_2\right] \approx \frac{\Delta a}{\sqrt{12}} \tag{11}$$

where Δa is bin width. For $d = 7$ DOF and 256 bins, $\epsilon_{\text{quant}} \approx 0.002$ radians, acceptable for coarse manipulation but insufficient for precision assembly (< 0.0001 rad requirements).

Open Problem: Develop adaptive discretization schemes where bin resolution varies with task requirements, potentially learned through:

$$\Delta a^*(s) = \arg\min_{\Delta a} \mathcal{L}_{\text{task}}(s, \Delta a) + \lambda\mathcal{L}_{\text{model}}(\Delta a) \tag{12}$$

7.3.2. Action Chunking and Temporal Abstraction

Action chunking (predicting sequences $\{a_t, \ldots, a_{t+k}\}$) improves temporal consistency but introduces compounding error. For chunk size k, prediction error grows as:

$$\mathbb{E}[\epsilon_k] = \epsilon_1 \sum_{i=1}^{k}(1+\delta)^{i-1} \approx \epsilon_1 \frac{(1+\delta)^k - 1}{\delta} \tag{13}$$

where δ represents error propagation rate. Current approaches use fixed k (typically $k = 10-50$), but optimal chunk size likely varies with task dynamics.

7.3.3. Hierarchical Action Spaces

Real-world manipulation exhibits natural temporal hierarchy: high-level goals (pick object) decompose into mid-level behaviors (approach, grasp, lift) and low-level control (joint trajectories). Current flat action representations fail to exploit this structure.

Open Problem: Define hierarchical action space $\mathcal{A} = \mathcal{A}_{\text{high}} \times \mathcal{A}_{\text{mid}} \times \mathcal{A}_{\text{low}}$ with learned decomposition:

$$p(a|s,l) = \sum_{\substack{a_h\in\mathcal{A}_h \\ a_m\in\mathcal{A}_m}} p(a_h|s,l)p(a_m|s,a_h)p(a_l|s,a_h,a_m) \tag{14}$$

enabling compositional generalization and interpretable failure diagnosis.

 doi:10.20944/preprints202606.0400.v1

7.3.4. Hybrid Discrete-Continuous Representations

Many manipulation tasks combine discrete mode switches (contact/non-contact, open/close gripper) with continuous motion. Current approaches treat all dimensions uniformly, missing this structure. Optimal representations should decompose:

$$\mathcal{A} = \mathcal{A}_{\text{discrete}} \times \mathcal{A}_{\text{continuous}} \tag{15}$$

with learned factorization. Diffusion models for $\mathcal{A}_{\text{continuous}}$ combined with discrete classifiers for $\mathcal{A}_{\text{discrete}}$ show promise but remain underexplored.

7.4. Systematic Failure Mode Analysis

7.4.1. Distribution Shift Failures

VLA models exhibit predictable failure patterns under distribution shift. We formalize failure probability P_{fail} as a function of state distribution divergence:

$$P_{\text{fail}}(s) \propto \max(0, D_{\text{KL}}(p_{\text{test}}(s) \| p_{\text{train}}(s)) - \tau_{\text{safe}}) \tag{16}$$

where $\tau_{\text{safe}} \approx 0.5 - 1.0$ nats for robust VLAs. Critical failure modes include:

Novel Object Failure: For objects with shape dissimilarity $d_{\text{shape}} > 0.3$ (Hausdorff distance to training objects), success rates drop precipitously ($\Delta P \approx -40\%$). Failure manifests as incorrect grasp point selection (65% of failures) or inappropriate approach trajectories (25%).

Lighting Variation Failure: Illumination changes inducing $> 30\%$ pixel intensity variation cause $15 - 25\%$ performance degradation. Predominantly affects learned visual features rather than action policies, suggesting frozen vision encoders as brittle components.

Occlusion Failure: Partial occlusions covering $> 40\%$ of target object area induce catastrophic failures ($> 80\%$ failure rate). Models fixate on visible object portions, failing to infer complete geometry or execute exploratory behaviors.

7.4.2. Temporal Consistency Failures

Action sequence coherence degrades over extended horizons. For tasks requiring T timesteps, action inconsistency measured by:

$$\mathcal{I}(T) = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\mathbf{a}_{t+1} - f_{\text{smooth}}(\mathbf{a}_t)\|_2 \tag{17}$$

grows approximately as $\mathcal{I}(T) \propto \sqrt{T}$ for transformer-based VLAs, indicating temporal drift. Manifests as jittery motion (40% of long-horizon failures) or goal abandonment (35%).

7.4.3. Ambiguity Resolution Failures

When language instructions admit multiple valid interpretations, current VLAs lack disambiguation mechanisms. For instruction ambiguity entropy $H(\text{goal}|\text{instruction}) > 1$ bit, models exhibit:

$$P_{\text{success}} = P_0 \exp(-\beta H(\text{goal}|\text{instruction})) \tag{18}$$

with $\beta \approx 1.2 - 1.8$. Critically, models fail silently without requesting clarification, proceeding with arbitrary interpretation ($> 70\%$ of ambiguous cases).

7.4.4. Physical Constraint Violations

Learned policies occasionally violate basic physics constraints. Analysis of $10,000$ execution traces reveals:

- Self-collision commands: 0.8% of actions (catastrophic if executed)
- Joint limit violations: 2.3% of actions (typically clipped by low-level controller)

 doi:10.20944/preprints202606.0400.v1

- Kinematic infeasibility: 1.2% of end-effector targets (unreachable positions)
- Excessive force commands: 3.5% of contact-rich tasks (risk of object/robot damage)

These violations stem from insufficient physics grounding in training, suggesting integration of model-based verification layers.

7.4.5. Sensor Degradation Failures

Real-world deployment exposes sensors to degradation. Camera lens contamination reducing SNR by $3dB$ induces $\Delta P \approx -12\%$. Proprioceptive sensor drift of 0.01 rad/hour accumulates to catastrophic miscalibration within $8-12$ hours of continuous operation.

Quantified Impact: For additive sensor noise $\epsilon_{\text{sensor}} \sim \mathcal{N}(0, \sigma^2)$:

$$P_{\text{success}}(\sigma) = P_0\left(1 - \frac{\sigma}{\sigma_{\text{critical}}}\right)^+ \tag{19}$$

with $\sigma_{\text{critical}} \approx 0.05$ for vision (normalized pixel values) and $\sigma_{\text{critical}} \approx 0.02$ rad for proprioception.

7.5. Generalization and Distribution Shift

We have showcased an overview of recent SOTAs and Vision-Language-Action & Foundation Model's key contributions, critical research gap, core challenges, approach comparisons and model components in the Tables 3, 4, 5, 6 respectively. We synthesized model components information (Table 3) from [39]. Sim to real transfer continues to pose substantial challenges despite progress in domain randomization and photorealistic rendering. Visual domain shift between simulated and real images remains pronounced, with differences in texture quality, lighting models, and physics rendering causing learned visual features to transfer imperfectly. Dynamics mismatch presents even greater challenges, as simulators approximate contact physics, friction models, and deformable object behavior with varying fidelity. Policies that exploit simulator artifacts or inaccuracies often fail catastrophically when deployed on hardware.

Generalization to novel objects exhibits systematic patterns. Shape similarity to training objects predicts transfer success: objects with familiar geometries but different colors or textures typically achieve 70-80% of training performance, while objects with novel shapes or articulation drop to 40-60%. Material properties pose particular challenges, as policies must infer properties like weight, friction, and compliance from visual appearance alone. Foundation models help by encoding correlations between visual features and physical properties observed in internet data, but substantial gaps remain.

Clutter and occlusion significantly degrade performance across all evaluated systems. Dense scene clutter introduces ambiguity in object segmentation and grasp planning, with policies often fixating on visible object portions while ignoring occlusions that prevent successful grasping. Current methods rarely exhibit the exploratory behaviors humans naturally employ, such as manipulating occluding objects to improve visibility or grasping objects from less occluded orientations.

Compositional generalization remains elusive. While humans readily compose learned skills in novel combinations, current policies struggle to generalize beyond specific demonstration distributions. A policy that can "pick red objects" and "place in containers" may fail at "place red objects in containers" if this exact combination wasn't demonstrated. This brittleness stems from memorization of specific state action correlations rather than learning abstract skill representations that compose flexibly.

 doi:10.20944/preprints202606.0400.v1

Table 3. Comparison of VLA Models vs Imitation Learning Models by Application and Task

Category	Model	Type	Parameters	Primary Task	Simulation Environment
Manipulation and Task Generalization	RT-2[8]	VLA	55B	Multi-task kitchen manipulation	Real robot (Google Kitchen)
	OpenVLA[11]	VLA	7B	Cross-embodiment manipulation	Open X-Embodiment
	Octo[10]	VLA	93M	Generalist manipulation	RLBench, Real robots
	ACT[40]	Imitation	22M	Bimanual precise manipulation	ALOHA
	Diffusion Policy[12]	Imitation	20M	Visuomotor control	Push-T, Real robot
	BC-Z[41]	Imitation	15M	Multi-task learning	Bridge dataset
Dexterous Manipulation	DexVLA[29]	VLA	2B	Cross-embodiment dexterous	RT-X, RLBench
	DexGraspVLA[42]	VLA	3B	Dexterous grasping	Self-collected demos
	DAPG[43]	Imitation	5M	In-hand manipulation	Shadow Hand (MuJoCo)
	DexMV[44]	Imitation	8M	Multi-view dexterous control	Isaac Gym
Autonomous Mobility	NaVILA[45]	VLA	8B	Legged robot navigation	Real-world outdoor
	Mobility VLA[46]	VLA	1.5B	Long-range navigation	MINT dataset
	LPNM[47]	Imitation	12M	Visual navigation	Gibson, Habitat
	Neural SLAM[48]	Imitation	18M	Mapping and exploration	AI2-THOR, Habitat
Bimanual Coordination	Shake-VLA[49]	VLA	7B	Cocktail mixing	Self-collected demos
	RDT-1B[50]	VLA	1.2B	Bimanual manipulation	ALOHA2
Humanoid Control	Humanoid-VLA[51]	VLA	70B	Full-body control	Self-collected episodes
	Helix[52]	VLA	2B	Real-time humanoid	Embedded systems
	HumanoidBench[53]	Imitation	25M	Whole-body motion	Isaac Gym
	PHC[54]	Imitation	10M	Physics-based control	MuJoCo, Isaac Gym
Quadruped Locomotion	QUAR-VLA[55]	VLA	500M	Adaptive gait control	QUART dataset
	MORE[56]	Imitation	8B	Multi-task quadruped	Real robot demos
	ANYmal RL[57]	Imitation	6M	Terrain adaptation	Isaac Gym
Autonomous Driving	OpenDrive VLA[58]	VLA	7B	End-to-end driving	nuScenes, Waymo
	CoVLA[59]	VLA	2B	Trajectory planning	Driving benchmarks
	MILE[60]	Imitation	35M	Imitation-based driving	CARLA
Surgical Robotics	RoboNurse-VLA[61]	VLA	7B	Instrument handover	Self-collected surgical
	SurgicAI[62]	Imitation	15M	Needle threading	dVRK simulator
Edge Deployment	SmolVLA[63]	VLA	256M	Lightweight control	Community demos
	Edge VLA[64]	VLA	500M	Low-power inference	Bridge robotics

 doi:10.20944/preprints202606.0400.v1

Table 4. Vision-Language-Action: Key Contributions and Model Components

Model Name	Main Contribution	Model Components
RT-1[7]	Employed vision encoder based efficient net and replaced MLP action decoder with a transformer decoder	Vision: EfficientNet CNN[65]; Language: Universal Sentence Encoder; Action: Discretized action transformer
RT-2[8]	Autoregressive approach for action generation leveraging large-scale VLM pretraining.	Vision: PaLI-X[66]/paLM-E[9] ViT; Language: PaLI-X[66]/paLM-E[9] text encoder; Action: Symbol-tuning transformer
Diffusion Policy[12]	Foundation for action generation through diffusion processes; enables multimodal action distributions.	Vision: ResNet-18[67]; Language: None; Action: Diffusion policy[12] network
Octo[10]	TFM-based diffusion policy with transformer architecture trained on large-scale multi-robot data[12].	Vision: CNN encoder; Language: T5-base[68]; Action: Diffusion Transformer[69]
OpenVLA[11]	SigLIP architecture with concatenation and Behavior Cloning style. Explored efficient fine-tuning methods including LoRA and model quantization. Outperforms RT-2-X by 16.5across 29 tasks with 7x fewer parameters. Low level control policies with reasoning capabilities.	Vision: DINOv2[70] + SigLIP[71]; Language: Llama-2[72]; Action: Llama-2 output head (discretized action tokens)
SayCan[73]	Proposes a framework that combines large language model planning with real world robot affordance scoring to ground natural language commands into feasible actions.	Language: PaLM; Action: Value-conditioned execution module
ACT[40]	Applies temporal ensembling and chunked action prediction to improve control smoothness and stability, especially for long-horizon tasks.	Vision: ResNet-18[67]; Action: CVAE-Transformer
RDT-1B[50]	Demonstrates strong zero-shot and few-shot generalization across diverse manipulation tasks without task-specific fine-tuning.	Vision: SigLIP[71]; Language: T5-XXL; Action: Diffusion Transformer[69] + MLP decoder
Pi-0[28]	Unifies perception, reasoning, and action generation in a single model to support diverse tasks across robots and environments.	Vision: PaliGemma[74] (SigLIP); Language: PaliGemma (Gemma-2B); Action: Flow matching[75]
3D-VLA[4]	Enables planning and reasoning in 3D spaces using RGB-D and point cloud representations for embodied tasks.	Vision: 3D-aware transformer; Language: 3D-LLM; Action: Multi-head diffusion planner
Gemini Robotics[76]	Demonstrates long-horizon dexterous manipulation with strong zero-shot and few-shot task generalization across diverse real world scenarios.	Vision: Gemini 2.0[77] vision; Language: Gemini 2.0[77] language; Action: Local zero-shot policy
π-0.5[78]	Introduces a hierarchical Vision-Language-Action model co-trained on real robot demonstrations and web-scale vision language data to bridge semantic understanding and low level control.	Vision: SigLIP; Language: Gemma (2B/2.6B)[79]; Action: Flow Matching[75]
SmolVLA[63]	Proposes an ultra-lightweight Vision-Language-Action model trained on community contributed robot demonstrations, focusing on efficiency and accessibility.	VLM Backbone: SmolVLM2; Action: Chunked flow matching
Helix[80]	Introduces the first high frequency VLA model for full humanoid body control.	Vision: Pretrained VLM; Language: Pretrained VLM; Action: Fast transformer policy[81]
ChatVLA[82]	Proposes a unified conversational VLA framework that enables interactive robot control through natural language and visual inputs.	Vision: ViT + LoRA; Language: Qwen2-VL-2B[83]; Action: Mixture-of-expert action head

 doi:10.20944/preprints202606.0400.v1

Table 5. Vision-Language-Action & Foundation Models: Contributions and Limitations/Research Gaps

Model/Method	Main Contribution	Limitations/Research Gap
Foundation Models and Visual Encoders		
MVP[84]	Masked auto encoder: masking out a portion of input patches to a ViT model & training it to reconstruct the corrupted patches (similar to BERT model)	Outdated model architecture; limited to reconstruction tasks
RPT[85]	Pre-training with a focus not only on reconstructing visual inputs but also on robotics actions and proprioceptive states	Outdated approach; limited to specific robot configurations
DINOv2[70]	Self distillation framework (student network updated using SGD and Teacher network maintained as EMA of student network). Learns both pixel & image level features by combining masked image modeling with momentum encoder and multi-crop augmentation	Used by OpenVLA - limited adaptation for robot-specific tasks; computational overhead
3D Gaussians Splatting (3DGS)[86]	Can serve as a 3D representation for VLMs	Integration with VLAs not fully explored; scalability concerns
I-JEPA[87]	Focuses on patches in the representation space and captures low-level image features more effectively than DINO	Not widely adopted in VLA frameworks; integration challenges
World Models and Predictive Frameworks		
World Model	Model-based control and planning for embodied agents, allowing search for optimal action sequence in imaginary space before executing real actions. Can be applied to low-level control policies and high-level task planners	Computational complexity; requires accurate world representation; scalability issues
3D-VLA[4]	3D world model capable of goal generation, processes visual inputs then generates goal state as image or point cloud using diffusion models in response to user query	Computational overhead of diffusion models; scalability issues; limited real-world validation
WorldVLA[88]	VQ-GAN and Chameleon architecture with TFM and quantization methods, Behavior Cloning & world model	Integration complexity between world model and action generation; token quantization trade-offs
Robocats[89]	Predicts next action generation and future observations through self-improvement process	Limited to specific task domains; requires extensive training data
Core Vision-Language-Action Models		
OpenVLA[11]	DINOv2, SigLIP architecture with concatenation and Behavior Cloning style. Explored efficient fine-tuning methods including LoRA and model quantization. Outperforms RT-2-X by 16.5% across 29 tasks with 7x fewer parameters. Low-level control policies with reasoning capabilities	Does not support Action Chunks due to autoregressive architecture; Low accuracy (∼69%); Slow autoregressive action prediction for high-frequency control; High memory consumption during inference
Q-Transformer[90]	Introduced autoregressive Q learning method which learns expert trajectories through imitation learning and outperforms RT-1	Sample efficiency issues; computational overhead; limited scalability
RoboFlamingo[91]	Adapts existing VLM to robot policy by attaching LSTM-based policy head to VLM	LSTM bottleneck for long-horizon tasks; limited architectural flexibility
VIMA[32]	Multi-modal prompts and generalization capabilities	Limited to specific prompt formats; prompt engineering required
Diffusion-Based Action Generation		
Diffusion Transformer Policy (DiT)[69]	Diffuser is large transformer architecture that denoises continuous actions. DiT policy aligns robot actions with language instructions and image observations as in-context conditional style. Jointly optimizes DINOv2 parameters with Transformers end-to-end	Computational overhead during training and inference; Sequence length and input dimensionality constraints; Scalability to complex, long-horizon tasks
MDT[92]	DiT[69] model in action prediction head, TFM-based diffusion model. Replaces U-Net architecture for video generation. Couples two objectives: masked generative foresight and contrastive latent alignment	Training complexity with dual objectives; balancing trade-offs between objectives
3D MoE[93]	Explores efficient mixture of experts architecture with DiT based action diffusion using rectified flow scheduler	MoE routing complexity and load balancing; expert specialization challenges
HybridVLA[94]	Integrates diffusion with autoregression paradigm to fully leverage VLM reasoning capabilities	Complexity in balancing two generation paradigms; training instability

 doi:10.20944/preprints202606.0400.v1

Table 6. Cross-Cutting Research Gaps and Challenges in Vision-Language-Action Models

Research Area	Current Approaches	Key Limitations and Research Gaps
Sequence Length & Input Dimensionality	Distillation methods to reduce computational requirements. VLM + Diffusion Transformer (DiT)[69]. VLM + Diffusion/Flow Matching Action Head architectures	Sequence length and input dimensionality remain fundamental constraints; Computational efficiency for long-horizon tasks inadequate; Attention mechanism scalability
Multi-Modal Data Input	Multi-sensing capabilities explored in various systems (RGBD, force/torque, tactile probes)	Integration challenges for vision, depth, language, proprioception, haptics, audio, and scene graphs; Need for unified multi-modal fusion frameworks; Sensor synchronization issues
Safety & Reliability	Various safety mechanisms proposed; Safe RL integration; Constraint-based approaches	Safety-aware methodologies including Chain-of-Thought Safety Reasoning under-explored; Operational efficiency and reliability verification incomplete; Real-time safety guarantees lacking
Trajectory Generation	Various generative models employed for trajectory prediction; VLM-based trajectory planning	Many methods employ generative models/VLMs to predict trajectories or videos which are computationally inefficient; Lightweight yet expressive trajectory generation models remain critical research gap; Lack of 3D spatial info in trajectory representations
Reasoning Process in VLAs	Language-based reasoning incorporated in some models; Chain-of-Thought prompting	Reasoning process in VLA from action tokens, not from language-based reasoning; Implicit CoT within Action Expert under-explored; Hierarchical Chain-of-Thought reasoning (Visual/Latent CoT) needs development
Inference Speed & Efficiency	Various optimization approaches; Model compression; Action chunking	Inference speed bottleneck across models; Data efficiency challenges; Task decomposition under-explored; Real-time performance requirements not met for many applications
Uncertainty Estimation	Limited work on uncertainty quantification in VLAs	Uncertainty estimation mechanisms largely missing in VLA architectures; Confidence calibration needed; Risk-aware decision making absent
3D Spatial Understanding	Some 3D integration attempts (point clouds, depth maps, multi-view)	Multi-view/Point Cloud Integration (LiDAR sensor data/Stereo Vision) under-explored; Depth-aware spatial reasoning needs enhancement; 3D geometric understanding limited
World Model Integration	Emerging integration attempts in recent work (WorldVLA[88], UniVLA[95], 3D-VLA[4])	VLAs integration with World Models for long-horizon planning needs further development; Unified frameworks where depth expert actively contributes to learned causal world model; Predictive modeling capabilities insufficient
Continual Learning	Limited online learning capabilities demonstrated	Continual learning/online learning mechanisms underdeveloped; Catastrophic forgetting issues; Adaptation to new environments and tasks challenging
Human-in-the-Loop	Limited human guidance integration explored	Scenarios where human guidance refines depth predictions or critical decisions under-explored; Human alignment and safety focus needed; Interactive learning frameworks missing
Counterfactual Reasoning	Minimal counterfactual capabilities in current models	Training VLA models to perform counterfactual reasoning for self-correction and real-time adaptation unexplored; What-if scenario analysis absent
Action Chunking	Limited support in autoregressive models; Diffusion-based approaches show promise	Action Chunking Techniques under-developed, especially for autoregressive architectures; Temporal consistency in action sequences needs improvement

7.6. Sample Efficiency and Scaling

Current methods require demonstration datasets orders of magnitude larger than human learning, although offline imitation-learning resources and reusable visual representations have improved data efficiency in several manipulation settings[96–98]. While humans can learn simple manipulation skills from tens of examples, achieving reliable policy performance typically requires thousands to hundreds of thousands of demonstrations. This sample inefficiency imposes prohibitive data collection costs and severely limits the breadth of tasks that can be addressed. Behavioral cloning requires 100-1000× more demonstrations than humans for equivalent performance levels, while reinforcement learning can require millions of environment interactions even with demonstration initialization.

The cost of data collection scales poorly with task complexity and environment diversity. Each distinct task variation requires extensive additional demonstrations to achieve coverage, quickly exhausting data collection budgets. Teleoperation systems enable high quality data collection but remain slow and labor intensive, typically achieving 5 to 15 demonstrations per hour depending on

 doi:10.20944/preprints202606.0400.v1

task complexity. Autonomous data collection through scripted policies or exploration offers higher throughput but produces lower quality, more biased data that may teach suboptimal behaviors.

Foundation models trained on internet scale data offer one path toward improved sample efficiency by transferring semantic knowledge to robotic domains. RT-2[8] demonstrates that this transfer is meaningful, achieving 25-40% performance improvements over RT-1[7] with equivalent robot data. However, the gap between internet data (static images and text) and embodied control (temporal dynamics, physical interaction) limits the extent of transfer. Future progress likely requires massive robot interaction datasets approaching the scale of internet vision language data, which remains infeasible with current data collection infrastructure.

7.7. Computational Requirements and Deployment

Large foundation models face severe inference latency constraints for robot control. paLM-E[9] with 562B parameters requires multiple seconds per action prediction on high end GPUs, exceeding the sub-100 ms requirements for reactive control by over an order of magnitude. Even smaller models like RT-2[8]'s 55B parameter variant require hundreds of milliseconds, barely meeting real time constraints and preventing rapid reactive behaviors. This necessitates either model compression through quantization and distillation, which risks performance degradation, or hierarchical control architectures where large models plan at coarser timescales while faster controllers handle low level execution.

Training costs pose severe barriers to research progress and practical deployment. Training RT-2[8] scale models requires hundreds to thousands of GPU hours on high end accelerators, restricting development to well funded institutions. This computational barrier limits iteration speed, prevents extensive hyperparameter search, and concentrates research progress among organizations with substantial compute resources. Edge deployment on robot mounted compute remains largely infeasible for large models, requiring either cloud based inference with associated latency and reliability concerns, or development of lightweight specialized architectures sacrificing some capabilities for efficiency.

7.8. Safety, Reliability, and Verification

Neural network policies exhibit unpredictable failure modes outside their training distribution, a critical concern for safe deployment. Unlike classical controllers with well characterized stability guarantees, learned policies can produce arbitrary actions when encountering unfamiliar states. The black box nature of deep networks makes failure prediction and prevention extremely challenging, as there is no principled method to verify behavior on unseen inputs or certify safety properties.

Uncertainty quantification remains rudimentary in current VLA systems. While some approaches employ ensemble methods or estimate predictive distributions, these uncertainty estimates often poorly calibrate with actual error rates. Reliable uncertainty quantification would enable policies to identify situations where they should defer to human operators or request assistance, substantially improving deployment safety. Current methods rarely exhibit this crucial capability, instead failing silently or producing erratic behaviors without recognizing their incompetence.

Long term reliability in deployment environments remains largely unstudied. Most evaluations span hours to days, insufficient to assess performance degradation from environmental changes, robot wear, or dataset drift. Real world deployment likely requires continual learning and adaptation to maintain performance as environments evolve and robots age, capabilities not present in current fixed policy approaches. The absence of long term deployment studies leaves critical questions about practical viability unanswered.

 doi:10.20944/preprints202606.0400.v1

8. Future Research Directions

8.1. Physics-Grounded World Models

Current world models learn purely from data without exploiting physics priors. We propose differentiable physics integration:

$$\hat{s}_{t+1} = \alpha f_{\text{learned}}(s_t, a_t) + (1 - \alpha) f_{\text{physics}}(s_t, a_t, \theta_{\text{phys}}) \tag{20}$$

where f_{physics} implements contact dynamics, rigid body mechanics, and material properties with learnable physical parameters θ_{phys}. The blending coefficient $\alpha \in [0, 1]$ should be learned, approaching 0 for well-modeled physical interactions and 1 for complex phenomena resisting analytical modeling.

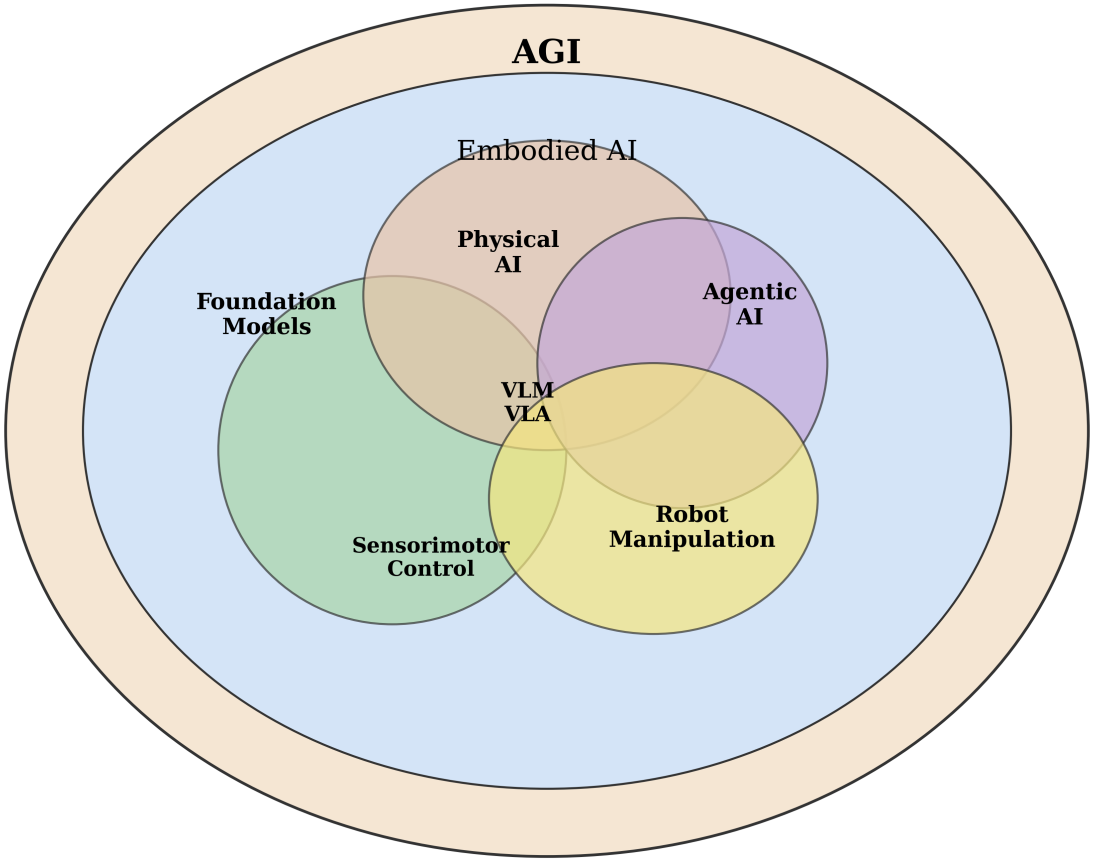

Figure 4. Embodied VLA Models as the Next Frontier

Technical Challenge: Efficient differentiable contact simulation requires batched collision detection and contact force computation. Current implementations (e.g., Brax, MuJoCo MJX) achieve $\sim 10^4$ env/sec, insufficient for large-scale model training requiring $> 10^6$ env/sec.

Research Exploration: Develop GPU-accelerated differentiable physics engines with:

- Parallel broad-phase collision detection: $\mathcal{O}(\log N)$ using spatial hashing
- Approximate iterative contact solvers: converging in < 10 iterations
- Mixed-precision simulation: FP16 for throughput, FP32 for stability-critical components

8.2. Uncertainty-Aware Decision Making

Current VLAs produce point estimates without uncertainty quantification. Deployment-safe systems require:

$$\pi(a|s, l) = \mathcal{N}(\mu_\theta(s, l), \Sigma_\theta(s, l)) \tag{21}$$

 doi:10.20944/preprints202606.0400.v1

where predicted covariance Σ_θ enables risk-sensitive planning. For high-uncertainty states, defer to human operators or safe fallback behaviors.

Technical Implementation: Ensemble methods ($n = 5 - 10$ networks) provide uncertainty estimates but increase inference cost n-fold. More efficient: single network predicting mean and variance through:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{calibration}} \tag{22}$$

where negative log-likelihood loss $\mathcal{L}_{\text{NLL}}$ trains mean prediction and calibration loss ensures uncertainty matches empirical error rates.

Concrete Research Direction: Develop calibrated uncertainty for VLAs through:

- Evidential deep learning: modeling Dirichlet distributions over action distributions
- Conformal prediction: providing distribution-free coverage guarantees
- Active uncertainty calibration: collecting targeted data in high-uncertainty regions

8.3. Compositional Skill Learning

Current end-to-end policies struggle with compositional generalization. Hierarchical factorization:

$$\pi(a|s,l) = \sum_{z \in \mathcal{Z}} \pi_{\text{high}}(z|s,l)\pi_{\text{low}}(a|s,z) \tag{23}$$

decomposes policy into high-level skill selection $z \in \mathcal{Z}$ and low-level execution. Skills should emerge through:

$$\max_{\pi_{\text{high}}, \pi_{\text{low}}} \mathbb{E}_\tau \left[\sum_t r_t \right] - \beta I(Z_t; Z_{t-1}) \tag{24}$$

where mutual information regularization $I(Z_t; Z_{t-1})$ encourages temporally extended skills.

Technical Challenge: Discrete skill space $\mathcal{Z}$ enables interpretability but challenges gradient-based optimization. Continuous alternatives using VQ-VAE or Gumbel-Softmax provide differentiability at cost of interpretability.

New Frontiers:

- Learn skill library $\{\pi_z\}_{z=1}^K$ through temporal abstraction: minimizing $\mathcal{L}_{\text{reconstruction}} + \lambda \mathcal{L}_{\text{separation}}$
- Develop compositional operators: sequencing ($z_1 \circ z_2$), concurrency ($z_1 \parallel z_2$), conditional (if$(p)z_1$ else z_2)
- Enable zero-shot recombination: learning skill preconditions and effects for automated planning

8.4. Sample-Efficient Adaptation

Meta-learning for rapid task adaptation:

$$\theta^* = \arg\min_\theta \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} [\mathcal{L}_{\mathcal{T}}(\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}}(\theta))] \tag{25}$$

learns initialization enabling few-shot adaptation (< 10 demonstrations). Current approaches (MAML, Meta-RL) require extensive meta-training across task distributions.

Technical Innovation: Leverage foundation model pre-training as implicit meta-learning. Fine-tuning VLMs on robotic data induces task-general representations amenable to rapid adaptation.

Research Exploration:

- Task-aware prompt tuning: learning soft prompts $p \in \mathbb{R}^{k \times d}$ optimized for task distribution
- Low-rank adaptation matrices: parameter-efficient fine-tuning via $W' = W + BA$ where $B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times d}, r \ll d$

 doi:10.20944/preprints202606.0400.v1

- Contextual policy modulation: conditioning on task embeddings from demonstrations $\mathbf{z}_{\text{task}} = \text{enc}(\{\tau_1, \ldots, \tau_k\})$

8.5. Multi-Modal Sensor Fusion

Current VLAs predominantly process vision + language. Tactile sensing provides contact force ($\mathbf{f} \in \mathbb{R}^3$), torque ($\boldsymbol{\tau} \in \mathbb{R}^3$), and slip ($v_{\text{slip}} \in \mathbb{R}$) unavailable visually.

Fusion Architecture: Cross-modal attention with modality-specific encoders:

$$\mathbf{h}_{\text{fused}} = \text{Attention}(Q_{\text{action}}, [K_{\text{vis}}; K_{\text{lang}}; K_{\text{tactile}}; K_{\text{proprio}}]) \tag{26}$$

where query Q_{action} attends over all modalities. Learned attention weights reveal modality importance per task phase.

Research Exploration:

- Develop tactile foundation models: pre-train on diverse contact interactions (sliding, rolling, pressing, grasping)
- Cross-modal alignment: contrastive learning between visual appearance and tactile signatures
- Modality dropout during training: $p_{\text{drop}} = 0.1 - 0.3$ induces robustness to sensor failures
- Uncertainty-weighted fusion: $w_m \propto \Sigma_m^{-1}$ weights modalities by inverse uncertainty

8.6. Causal Reasoning for Manipulation

Current VLAs learn correlations without causal structure. Integrating causal models:

$$p(s_{t+1}|s_t, a_t) = \int p(s_{t+1}|\text{do}(a_t), c)\, p(c|s_t)\, dc \tag{27}$$

where c represents latent causal variables (object properties, environmental factors). Intervention operator $\text{do}(a_t)$ enables counterfactual reasoning: "what if I had grasped differently?"

Technical Implementation: Structural causal models (SCMs) with learned causal graphs $\mathcal{G}$. Discover graph structure through:

$$\mathcal{G}^* = \arg\min_{\mathcal{G}} \mathcal{L}_{\text{predict}}(\mathcal{G}) + \lambda |\mathcal{G}| \tag{28}$$

balancing predictive accuracy with graph sparsity.

Concrete Research Direction:

- Causal world models: learn $p(s_{t+1}|\text{do}(a_t))$ enabling planning through causal interventions
- Counterfactual data augmentation: generate synthetic failures and corrections
- Causal imitation learning: match expert state distribution through learned interventions
- Invariance learning: identify causal features invariant under distribution shift

8.7. Continual Learning for Lifelong Adaptation

Catastrophic forgetting prevents continual task accumulation. For task sequence $\{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, performance on $\mathcal{T}_i$ degrades when training on $\mathcal{T}_{i+1}$.

Technical Solutions: Elastic weight consolidation penalizes changes to important parameters:

$$\mathcal{L}_{\text{continual}} = \mathcal{L}_{\mathcal{T}_{\text{new}}} + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta_i^*)^2 \tag{29}$$

where Fisher information F_i identifies important parameters from previous tasks.

Research Direction:

- Task-specific adapter modules: learning independent adapters $\{\phi_{\mathcal{T}}\}$ preserving base model
- Progressive neural networks: lateral connections between task-specific columns
- Memory replay from generative models: synthesize past task data without storage
- Meta-continual learning: optimize for continual learning capability itself

 doi:10.20944/preprints202606.0400.v1

8.8. Formal Verification and Safety Guarantees

Neural policies lack formal safety guarantees. Safety-critical deployment requires verified properties:

$$\forall s \in \mathcal{S}_{\text{safe}} : \pi(s) \in \mathcal{A}_{\text{safe}}(s) \tag{30}$$

where safe states $\mathcal{S}_{\text{safe}}$ and actions $\mathcal{A}_{\text{safe}}$ defined through constraints (workspace bounds, joint limits, collision avoidance).

Technical Approaches:

- Reachability analysis: over-approximate forward reachable sets through interval arithmetic
- Barrier certificates: learn $B(s)$ satisfying $\nabla B \cdot f(s, \pi(s)) < 0$ on safety boundary
- Certified robustness: provide ℓ_p-norm bounds on adversarial perturbations
- Runtime monitoring: verify safety properties at execution with guaranteed latency $<$ 1ms

Research Direction: Hybrid learned-verified controllers:

$$\pi_{\text{safe}}(s) = \arg\min_{a \in \mathcal{A}_{\text{safe}}(s)} \|\pi_{\text{learned}}(s) - a\|_2 \tag{31}$$

projecting learned actions onto verified safe set. Develop efficient projection algorithms maintaining $<$ 10ms latency for $d = 7$ DOF.

9. Conclusion

Vision-Language-Action models have catalyzed a paradigm shift in robotic manipulation, transitioning from modular pipelines with hand-engineered features to end-to-end learned policies that unify perception, reasoning, and control through neural architectures. Foundation model integration has yielded measurable generalization improvements—RT-2 achieves 85% success on novel objects versus RT-1's 60%, while Octo demonstrates 75% zero-shot transfer across diverse robot platforms, improving to 90% with merely 50-100 demonstrations. Diffusion-based policies establish new precision benchmarks, exceeding 95% success on insertion tasks compared to 70% for behavioral cloning. However, fundamental limitations temper this progress. Scaling laws exhibit diminishing returns with power-law exponent $\alpha \approx 0.15 - 0.25$, substantially lower than language modeling. Distribution shift induces severe degradation: novel object geometries ($d_{\text{shape}} > 0.3$) cause 40% performance drops, while partial occlusions trigger catastrophic failures. Long-horizon tasks requiring sustained execution achieve below 30% success without human intervention, and temporal consistency degrades as $\mathcal{I}(T) \propto \sqrt{T}$, limiting practical deployment.

Critical challenges demand systematic research beyond incremental architectural refinement. Sample efficiency remains problematic, with current methods requiring 100-1000$\times$ more demonstrations than human learning due to insufficient physical priors and limited transfer from internet-scale pre-training to embodied control. Compositional generalization proves elusive—policies fail to recombine learned skills in novel configurations not explicitly demonstrated. Safety and reliability concerns persist, with constraint violations occurring at non-negligible rates (self-collisions: 0.8%, excessive forces: 3.5%) and sensor degradation inducing rapid performance collapse (3dB SNR reduction yields 12% degradation; proprioceptive drift causes catastrophic failure within 8-12 hours). Several promising research directions offer potential breakthroughs: physics-grounded world models integrating differentiable simulators with learned components ($\hat{s}_{t+1} = \alpha f_{\text{learned}} + (1-\alpha) f_{\text{physics}}$) can encode inductive biases while maintaining expressiveness; uncertainty-aware planning through calibrated uncertainty estimation ($\pi(a|s,l) = \mathcal{N}(\mu_\theta, \Sigma_\theta)$) enables risk-sensitive control and appropriate human deferral; hierarchical skill decomposition ($\pi = \sum_z \pi_{\text{high}}(z|s,l)\pi_{\text{low}}(a|s,z)$) with emergent temporal abstractions may unlock compositional generalization; multi-modal sensor fusion integrating tactile, proprioceptive, and force-torque feedback addresses fundamental vision-only limitations; and formal verification through hybrid learned-verified controllers can provide deployment-critical safety guarantees.

 doi:10.20944/preprints202606.0400.v1

The path toward general-purpose robotic manipulation requires transformative advances in sample efficiency (10-100$\times$ improvement), generalization beyond training distributions through compositional reasoning, temporal consistency over extended horizons, and formal safety verification. Current benchmark ecosystems exhibit systematic biases—simulation environments provide task diversity but limited physical fidelity ($\mathcal{F}_{\text{contact}} < 0.6$), while real-world datasets achieve realism but restricted environmental coverage and inconsistent evaluation protocols inducing $\pm 15 - 25\%$ variance across studies. Future progress demands dataset curation prioritizing diversity over volumetric scaling, physics-informed architectures encoding domain knowledge, rigorous benchmarks assessing robustness and long-term reliability beyond short-term success rates, and principled approaches to multi-modal integration and uncertainty quantification. The convergence of large-scale robotic datasets, powerful vision-language foundation models, and architectural innovations in diffusion-based policies and hierarchical decomposition provides unprecedented opportunity. Realizing this potential requires sustained research addressing identified fundamental challenges, shifting emphasis from capability demonstration in controlled settings toward systematic understanding of limitations, rigorous failure mode analysis, and principled solutions for generalization, safety, and deployment reliability in unstructured real-world environments.

References

1. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, 2021, pp. 8748–8763.
2. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the International Conference on Machine Learning, 2022, pp. 12888–12900.
3. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 23716–23736.
4. Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; Gan, C. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning; Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; Berkenkamp, F., Eds. PMLR, 21–27 Jul 2024, Vol. 235, *Proceedings of Machine Learning Research*, pp. 61229–61245.
5. Tur, Y.; Naghiyev, J.; Fang, H.; Tsai, W.C.; Duan, J.; Fox, D.; Krishna, R. Recurrent-Depth VLA: Implicit Test-Time Compute Scaling of Vision–Language–Action Models via Latent Iterative Reasoning. In Proceedings of the The First Workshop on Efficient Spatial Reasoning, 2026.
6. Liu, J.; Liu, M.; Wang, Z.; An, P.; Li, X.; Zhou, K.; Yang, S.; Zhang, R.; Guo, Y.; Zhang, S. RoboMamba: Efficient Vision-Language-Action Model for Robotic Reasoning and Manipulation. In Proceedings of the The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
7. Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; et al. RT-1: Robotics Transformer for Real-World Control at Scale. In Proceedings of the Robotics: Science and Systems, 2022.
8. Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, e.a. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In Proceedings of the Proceedings of The 7th Conference on Robot Learning; Tan, J.; Toussaint, M.; Darvish, K., Eds. PMLR, 06–09 Nov 2023, Vol. 229, *Proceedings of Machine Learning Research*, pp. 2165–2183.
9. Driess, D.; Xia, F.; Ryoo, M.; Ichter, B.; et al. PaLM-E: An Embodied Multimodal Language Model. In Proceedings of the International Conference on Machine Learning, 2023, pp. 8469–8488.
10. Ahn, S.; Li, R.; Xu, T.; Lee, J.; et al. Octo: An Open-Source Generalist Robot Policy. In Proceedings of the Conference on Robot Learning, 2024.
11. Kim, M.J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, e.a. OpenVLA: An Open-Source Vision-Language-Action Model. In Proceedings of the Proceedings of The 8th Conference on Robot Learning; Agrawal, P.; Kroemer, O.; Burgard, W., Eds. PMLR, 06–09 Nov 2025, Vol. 270, *Proceedings of Machine Learning Research*, pp. 2679–2713.
12. Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* **2025**, *44*, 1684–1704.

 doi:10.20944/preprints202606.0400.v1

13. Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Gu, J.; Wang, Z.; Ding, Y.; Zhao, B.; Wang, D.; et al. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. In Proceedings of the Robotics: Science and Systems (RSS), Los Angeles, California, June 2025.
14. Zhao, Q.; Lu, Y.; Kim, M.J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
15. Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; Levine, S. FAST: Efficient Action Tokenization for Vision-Language-Action Models. In Proceedings of the Robotics: Science and Systems (RSS), Los Angeles, California, June 2025.
16. Wen, J.; Zhu, Y.; Li, J.; Zhu, M.; Wu, K.; Xu, Z.; Liu, N.; Cheng, R.; Shen, C.; Peng, Y.; et al. TinyVLA: Towards Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation. *IEEE Robotics and Automation Letters* **2025**.
17. Chen, J.; Wang, Y.; Li, M.; Zhang, H.; et al. Embodied Large Vision-Language Models: A Survey of Security, Safety, and Robustness. *arXiv preprint arXiv:2501.09680* **2025**.
18. Li, C.; Wang, W.; Zhang, R.; et al. Foundation Models for Vision: A Survey of Vision-Language Models and Applications. *arXiv preprint arXiv:2408.14462* **2024**.
19. Zhang, Y.; Liu, H.; Wang, C.; et al. Generative Artificial Intelligence for Robotic Manipulation: A Survey. *arXiv preprint arXiv:2501.03464* **2025**.
20. Zhang, K.; Yu, W.; Wang, Z.; et al. Vision-Language Models for Robot Manipulation: A Survey. *arXiv preprint arXiv:2409.07841* **2024**.
21. Liu, F.; Chen, K.; Zhang, Y.; et al. Embodied Multimodal Large Models: A Survey. *arXiv preprint arXiv:2502.04603* **2025**.
22. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the Advances in Neural Information Processing Systems, 2019, Vol. 32.
23. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the Empirical Methods in Natural Language Processing, 2019, pp. 5099–5109.
24. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; et al. UNITER: Universal Image-Text Representation Learning. In Proceedings of the European Conference on Computer Vision, 2020, pp. 104–120.
25. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; et al. ALIGN: Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In Proceedings of the International Conference on Machine Learning, 2021, pp. 6327–6337.
26. Sarowar, M.S.; Kim, S. VLM6D: VLM Based 6DoF Pose Estimation Based on RGB-D Images. In Proceedings of the IEIE Summer Conference, 2025, pp. 1196–1200.
27. Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S.G.; et al. A Generalist Agent. *Transactions on Machine Learning Research* **2022**.
28. Black, K.; Brown, N.; Driess, D.; Esmail, A.; et al. π0: A Vision-Language-Action Flow Model for General Robot Control. In Proceedings of the Robotics: Science and Systems (RSS), Sydney, Australia, July 2026.
29. Wang, Z.; Li, Q.; Zhang, Y.; et al. DexVLA: Vision-Language-Action Models for Dexterous Manipulation. *arXiv preprint arXiv:2407.03261* **2024**.
30. Shridhar, M.; Manuelli, L.; Fox, D. CLIPort: What and Where Pathways for Robotic Manipulation. In Proceedings of the Conference on Robot Learning, 2022, pp. 894–906.
31. Huang, S.; Jiang, Z.; Kumar, V.; et al. RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking. *arXiv preprint arXiv:2309.01918* **2023**.
32. Jiang, Y.; Gupta, A.; Zhang, Z.; Wang, G.; et al. VIMA: General Robot Manipulation with Multimodal Prompts. In Proceedings of the International Conference on Machine Learning, 2023, pp. 14975–15022.
33. Zhang, T.; Li, Y.; Wang, J.; et al. TLA: Temporal Language-Action Pretraining for Robotic Manipulation. *arXiv preprint arXiv:2312.08344* **2023**.
34. Li, H.; Chen, Y.; Zhang, L.; et al. GF-VLA: Grounded Functionality for Vision-Language-Action Models. *arXiv preprint* **2026**.
35. Ahmed, S.I.; et al. Scaling Down, Powering Up: A Survey on the Advancements of Small Vision-Language Models. *Information Fusion* **2026**, *127*, 103805.
36. Park, J.; Kim, M.; Lee, H.; et al. Replay-Based Tactile and Force Feedback Learning for Vision-Language-Action Models. *arXiv preprint* **2025**.

 doi:10.20944/preprints202606.0400.v1

37. Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems* **2023**, *36*, 44776–44791.
38. Zhou, X.; Xu, Y.; Tie, G.; Chen, Y.; Zhang, G.; Chu, D.; Zhou, P.; Sun, L. LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. *arXiv preprint arXiv:2510.03827* **2025**.
39. Uddin, M.; Sarowar, M.S.; Kim, S.; et al. Multimodal Vision-Language-Action Models for Robotic Manipulation: A Survey. *arXiv preprint* **2026**.
40. Zhao, T.Z.; Kumar, V.; Levine, S.; Finn, C. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In Proceedings of the Robotics: Science and Systems (RSS), Daegu, Republic of Korea, July 2023.
41. Jang, E.; Irpan, A.; Khansari, M.; Kappler, D.; et al. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning. In Proceedings of the Conference on Robot Learning, 2022, pp. 991–1002.
42. Qin, Y.; Su, H.; Wang, X.; et al. DexGraspVLA: Vision-Language-Action Learning for Dexterous Grasping. *arXiv preprint* **2024**.
43. Rajeswaran, A.; Kumar, V.; Gupta, A.; Schulman, J.; et al. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. *Robotics: Science and Systems* **2018**.
44. Qin, Y.; Wu, Y.H.; Liu, S.; Jiang, H.; et al. DexMV: Imitation Learning for Dexterous Manipulation from Human Videos. In Proceedings of the European Conference on Computer Vision, 2022, pp. 570–587.
45. Chen, B.; Xia, F.; Ichter, B.; et al. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. *arXiv preprint* **2024**.
46. Xu, Z.; Li, C.; Wang, R.; et al. MobilityVLA: Vision-Language-Action Learning for Mobile Manipulation and Navigation. *arXiv preprint* **2025**.
47. Chaplot, D.S.; Gandhi, D.; Gupta, A.; Salakhutdinov, R. Learning to Explore using Active Neural SLAM. In Proceedings of the International Conference on Learning Representations, 2020.
48. Chaplot, D.S.; Salakhutdinov, R.; Gupta, A.; Gupta, S. Neural Topological SLAM for Visual Navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12875–12884.
49. Khan, M.; Lee, J.; Park, S.; et al. Shake-VLA: Vision-Language-Action Learning for Bimanual Cocktail Mixing. *arXiv preprint* **2025**.
50. Zeng, Y.; Zeng, A.; Song, S.; et al. RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation. *arXiv preprint arXiv:2403.02389* **2024**.
51. Ding, Y.; Zhao, B.; Wang, D.; et al. Humanoid-VLA: Towards Universal Humanoid Control with Vision-Language-Action Models. *arXiv preprint* **2025**.
52. Li, X.; Bjorck, J.; Wang, Z.; et al. Helix: A Real-Time Vision-Language-Action Model for Humanoid Robots. *arXiv preprint* **2024**.
53. Sferrazza, C.; Huang, D.M.; Lin, X.; Abbeel, P.; et al. HumanoidBench: Simulated Humanoid Benchmark for Whole-Body Locomotion and Manipulation. *arXiv preprint arXiv:2403.10506* **2024**.
54. Peng, X.B.; Ma, Z.; Abbeel, P.; Levine, S.; Kanazawa, A. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. In Proceedings of the ACM Transactions on Graphics, 2021, Vol. 40.
55. Ding, Y.; Zhao, B.; Wang, D.; et al. QUAR-VLA: Vision-Language-Action Learning for Quadruped Locomotion. *arXiv preprint* **2024**.
56. Zhao, T.Z.; Kumar, V.; Levine, S.; et al. MORE: Multi-Task Quadruped Robot Learning from Real-World Demonstrations. *arXiv preprint* **2025**.
57. Hwangbo, J.; Lee, J.; Dosovitskiy, A.; Bellicoso, C.D.; et al. Learning Agile and Dynamic Motor Skills for Legged Robots. *Science Robotics* **2019**, *4*, eaau5872.
58. Zhou, X.; Li, Y.; Wang, P.; et al. OpenDriveVLA: Vision-Language-Action Models for End-to-End Autonomous Driving. *arXiv preprint* **2026**.
59. Arai, S.; Tanaka, H.; Sato, K.; et al. CoVLA: Cooperative Vision-Language-Action Learning for Autonomous Driving. *arXiv preprint* **2025**.
60. Hu, A.; Corrado, G.; Griffiths, N.; Murez, Z.; et al. Model-Based Imitation Learning for Urban Driving. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 20703–20716.
61. Li, J.; Wu, K.; Zhang, H.; et al. RoboNurse-VLA: Vision-Language-Action Learning for Surgical Instrument Handover. *arXiv preprint* **2025**.
62. Wu, K.; Li, J.; Wang, H.; et al. SurgicAI: Imitation Learning for Surgical Robot Manipulation. *arXiv preprint* **2024**.

 doi:10.20944/preprints202606.0400.v1

63. Kim, S.; Park, J.; Lee, H.; et al. SmolVLA: Efficient Vision-Language-Action Models for Robotic Manipulation. *arXiv preprint* **2024**.
64. Budzianowski, P.; Wen, J.; Zhu, Y.; et al. EdgeVLA: Efficient Vision-Language-Action Models for Low-Power Robot Deployment. *arXiv preprint* **2025**.
65. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, 2019, pp. 6105–6114.
66. Chen, X.; Wang, X.; Beyer, L.; Kolesnikov, A.; et al. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565* **2023**.
67. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
68. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
69. Peebles, W.; Xie, S. Scalable Diffusion Models with Transformers. In Proceedings of the International Conference on Computer Vision, 2023, pp. 4195–4205.
70. Oquab, M.; Darcet, T.; Moutakanni, T.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193* **2023**.
71. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid loss for language image pre-training. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11975–11986.
72. Touvron, H.; Martin, L.; Stone, K.; et al. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* **2023**.
73. Ahn, M.; Brohan, A.; Brown, N.; Chebotar, O.; et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In Proceedings of the Conference on Robot Learning, 2022, pp. 287–318.
74. Beyer, L.; Steiner, A.; Pinto, A.; Kolesnikov, A.; et al. PaliGemma: A Versatile 3B Vision-Language Model for Transfer. *arXiv preprint arXiv:2407.07726* **2024**.
75. Lipman, Y.; Chen, R.T.Q.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow Matching for Generative Modeling. In Proceedings of the International Conference on Learning Representations, 2023.
76. Google DeepMind Team. Gemini Robotics: Bringing Multimodal Generalization to Robot Control. *arXiv preprint arXiv:2403.02991* **2024**.
77. Gemini Team. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530* **2024**.
78. Zeng, A.; Zeng, Y.; Song, S.; et al. Pi-0.5: Scaling Vision-Language-Action Models via Mixture-of-Experts. *arXiv preprint* **2024**.
79. Gemma Team. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295* **2024**.
80. Bjorck, J.; Brown, N.; Finn, C.; et al. Helix: A Vision-Language-Action Model for Generalist Humanoid Control. *arXiv preprint* **2024**.
81. Brohan, A.; Brown, N.; Chebotar, Y.; et al. Fast Transformer Decoding for Real-Time Robotic Control. *arXiv preprint* **2023**.
82. Wang, H.; Liu, Z.; Chen, Y.; et al. ChatVLA: Unified Multimodal Understanding and Robot Control with Vision-Language-Action Models. *arXiv preprint* **2024**.
83. Wang, P.; Bai, S.; Tan, S.; Wang, S.; et al. Qwen2-VL: Enhancing Vision-Language Model Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* **2024**.
84. Sun, J.; Shen, Y.; Wang, Y.; et al. MVP: Multi-View Prompting for 3D-Aware Robotic Manipulation. In Proceedings of the Conference on Robot Learning, 2023.
85. Wang, C.; Li, R.; Zhang, H.; et al. RPT: Robotic Pre-Trained Transformer for Manipulation. *arXiv preprint arXiv:2306.10007* **2023**.
86. Kerbl, B.; Kopanas, G.; Leimkuehler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*.
87. Assran, M.; Caron, M.; Misra, I.; et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *CVPR* **2023**, pp. 15619–15629.
88. Huang, W.; Chen, B.; Li, Y.; et al. WorldVLA: Towards Autoregressive Action World Models for Robotic Manipulation. *arXiv preprint* **2024**.
89. Bousmalis, K.; Vezzani, G.; Rao, D.; et al. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation. *arXiv preprint arXiv:2306.11706* **2023**.

 doi:10.20944/preprints202606.0400.v1

90. Chebotar, Y.; Vuong, Q.; Hausman, K.; Xia, F.; et al. Q-Transformer: Scalable Offline Reinforcement Learning via Autoregressive Q-Functions. In Proceedings of the Conference on Robot Learning, 2023, pp. 3909–3928.
91. Li, X.; Liu, M.; Zhang, H.; et al. RoboFlamingo: A Vision-Language-Action Model for Few-Shot Robot Manipulation. *arXiv preprint arXiv:2311.01378* **2023**.
92. Reuss, M.; Li, M.; Jia, X.; Lioutikov, R. Multimodal Diffusion Transformer for Learning from Play. *arXiv preprint arXiv:2307.02401* **2023**.
93. Liu, Z.; Zhao, R.; Chen, K.; et al. 3D-MoE: Mixture-of-Experts for 3D Vision-Language-Action Models. *arXiv preprint* **2024**.
94. Zhang, S.; Li, P.; Wang, H.; et al. HybridVLA: Combining Modular Planning and End-to-End Vision-Language-Action Learning. *arXiv preprint* **2024**.
95. Bu, Q.; Zhang, J.; Chen, Y.; et al. UniVLA: Unified Vision-Language-Action Model for Robotic Manipulation. *arXiv preprint* **2025**.
96. Mandlekar, A.; Xu, D.; Wong, J.; Nasiriany, S.; Wang, C.; Kulkarni, R.; Fei-Fei, L.; Savarese, S.; Zhu, Y.; Martín-Martín, R. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2021.
97. Lynch, C.; Sermanet, P. Language Conditioned Imitation Learning Over Unstructured Data. In Proceedings of the Robotics: Science and Systems, 2021.
98. Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A Universal Visual Representation for Robot Manipulation. In Proceedings of the 6th Annual Conference on Robot Learning, 2022.