Vision-Language-Action models for pick-and-place tasks in embodied AI

Vision-Language-Action models achieve 50-75% success rates on pick-and-place tasks with strong generalization to novel objects and backgrounds, though they remain limited by inadequate spatial reasoning, difficulties with multi-step planning and cluttered environments, and substantial computational requirements that constrain real-world deployment.

Abstract

This systematic review of 10 sources examines Vision-Language-Action (VLA) models for pick-and-place tasks in embodied AI, revealing substantial progress alongside persistent limitations. Current VLA architectures achieve success rates ranging from approximately 50% on complex spatial tasks to over 70% on structured pick-and-place operations , with performance heavily dependent on task complexity and environmental conditions. Architectural innovations prove more impactful than raw model scale: 7B-parameter models employing componentized designs with diffusion action transformers or fused vision encoders outperform the 55B-parameter RT-2-X by 16-35% in absolute success rates . Models demonstrate strong generalization to novel objects and backgrounds , with RT-2 exhibiting emergent capabilities such as interpreting semantically indicated locations and performing rudimentary reasoning not present in training data .

However, significant limitations constrain current VLA capabilities for pick-and-place tasks. All models struggle with complex manipulation requiring multi-step planning , precise spatial reasoning , and operation in heavily occluded or cluttered environments . Current VLMs prioritize high-level semantic understanding while neglecting low-level spatial features critical for manipulation . Computational requirements remain substantial, with training demanding extensive GPU resources and inference speeds often inadequate for high-frequency control . Cross-robot transfer typically requires fine-tuning rather than zero-shot deployment . The evidence indicates that while VLA models represent a promising paradigm for generalizable robotic manipulation, achieving reliable performance on complex, real-world pick-and-place tasks requires continued advances in spatial reasoning, action representation, and sample-efficient training methodologies.

Paper search

We performed a semantic search using the query ”Vision-Language-Action models for pick-and-place tasks in embodied AI” across over 138 million academic papers from the Elicit search engine, which includes all of Semantic Scholar and OpenAlex.

We retrieved the 50 papers most relevant to the query.

Screening

We screened in sources based on their abstracts that met these criteria:

- **Multimodal VLA Integration**: Does this study involve Vision-Language-Action models or multimodal models that integrate visual perception, natural language processing, and action generation?
- **Manipulation Task Focus**: Is this research focused on pick-and-place tasks, object manipulation, or grasping tasks in robotic systems?
- **Embodied AI Systems**: Does this study involve embodied AI agents, robotic systems, or simulated environments where agents perform physical manipulation?

- **Performance Metrics**: Does this research report quantitative or qualitative performance metrics for manipulation tasks?
- **Beyond Computer Vision Only**: Does this study go beyond focusing solely on computer vision and include action generation or language understanding components?
- **Beyond Language Processing Only**: Does this research go beyond pure natural language processing or dialogue systems to include embodied interaction?
- **Manipulation vs Navigation Focus**: Does this study focus on manipulation tasks rather than exclusively on navigation, locomotion, or other non-manipulation tasks?
- **Empirical Evidence**: Does this study include empirical evaluation or technical implementation rather than being purely theoretical, opinion-based, or editorial content?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **VLA Model Architecture**:

  Extract comprehensive details about the vision-language-action model including:

- Base model or foundation model used (e.g., GPT-4o, PaLI, etc.)
- Specific architectural components (action prediction module, attention mechanisms, etc.)
- Model size (number of parameters)
- Novel architectural innovations or modifications
- How vision, language, and action components are integrated
- Action representation method (text tokens, continuous values, quantization, etc.)

- **Training Methodology**:

  Extract training approach details including:

- Training data types used (robotic trajectories, web data, simulation, real-world)
- Training procedure (co-fine-tuning, transfer learning, from-scratch, etc.)
- Data scale (number of trajectories, hours of data, etc.)
- Robot platforms used for data collection
- Pre-training vs fine-tuning stages
- Any data augmentation or synthesis techniques

- **Pick-Place Task Details**:

  Extract specific information about pick-and-place tasks including:

- Types of objects manipulated (everyday objects, novel objects, specific categories)
- Task complexity (single-step vs multi-step, planning requirements)
- Environmental settings (cluttered, clean, varied backgrounds)
- Language instruction types (natural language commands, structured prompts)
- Success criteria and task definitions
- Any task variations or difficulty levels tested

- **Performance Results**:

  Extract quantitative performance metrics including:

- Success rates for pick-and-place tasks (overall and by condition)
- Comparison with baseline methods or other VLA models
- Performance across different object types or environments
- Statistical significance and confidence intervals where reported
- Any learning curves or sample efficiency metrics
- Execution time or efficiency measures

- **Generalization Assessment**:

  Extract evidence of model generalization including:

- Performance on unseen objects or novel object categories
- Adaptation to new environments or backgrounds
- Cross-robot platform transfer capabilities
- Zero-shot vs few-shot performance comparisons
- Ability to handle instructions not seen during training
- Robustness to variations in lighting, viewpoint, or scene configuration

- **Evaluation Setup**:

  Extract experimental methodology details including:

- Robot hardware platforms used (Franka, UR5, etc.)
- Simulation environments if applicable
- Number of evaluation trials or experiments conducted
- Evaluation datasets or benchmark suites used
- Camera setup and sensory inputs
- Evaluation protocol and experimental controls

- **Critical Insights**:

  Extract key findings and insights including:

- What factors most impact pick-and-place success (architectural choices, training data, etc.)
- Identified failure modes or common error patterns
- Insights about scaling behaviors or model size effects
- Important design decisions that improved performance
- Comparison insights between different VLA approaches
- Any emergent capabilities or unexpected behaviors observed

- **Limitations**:

  Extract identified limitations and challenges including:

- Specific types of pick-and-place tasks that remain challenging
- Environmental conditions that cause failures
- Computational requirements or efficiency constraints
- Data requirements or sample efficiency issues
- Generalization boundaries or failure cases

- Technical limitations of current VLA approaches for manipulation

Characteristics of Included Studies

The reviewed literature encompasses a diverse range of Vision-Language-Action (VLA) model architectures, training methodologies, and evaluation approaches for pick-and-place tasks in embodied AI. Table 1 summarizes the key characteristics of each included source.

Study	Full Text Retrieved?	Study Type	Model(s) Examined	Base Architecture	Model Size
Qixiu Li et al., 2024	Yes	Primary study	CogACT	VLM with diffusion transformer	~7B parameters
Beichen Wang et al., 2024	Yes	Primary study	SeeDo	GPT-4o	Not mentioned
Kechun Xu et al., 2025	Yes	Primary study	A^2	MaskCLIP + GraspNet	Not mentioned
Moo Jin Kim et al., 2024	Yes	Primary study	OpenVLA	Llama 2 + DINOv2/SigLIP	7B parameters
P. Guruprasad et al., 2024	Yes	Benchmarking study	GPT-4o, OpenVLA, JAT	Multiple architectures	OpenVLA: 7B
Jianke Zhang et al., 2025	Yes	Primary study	UP-VLA	Phi-1.5 + CLIP-ViT	Not mentioned
Muhayy ud Din et al., 2025	Yes	Systematic review	Multiple VLA models	Transformer-based	Various
Lingling Fan et al., 2024	No (abstract only)	Primary study	VLA for grasping	Pre-trained VLM	Not mentioned
Jiange Yang et al., 2023	Yes	Primary study	TPM	Foundation models	Not mentioned
Anthony Brohan et al., 2023	Yes	Primary study	RT-2	PaLI-X and PaLM-E	Billions to tens of billions

The included studies represent both primary research developing novel VLA architectures and benchmarking/review studies that provide comparative analyses across multiple models. The majority of primary studies focus on models in the 7B parameter range, with RT-2 representing the largest model examined at tens of billions of parameters .

Model Architectures and Action Representations

VLA models for pick-and-place tasks employ diverse strategies for integrating vision, language, and action components. Table 2 presents the architectural approaches across studies.

Model	Vision Encoder	Language Model	Action Module	Action Representation
CogACT	VLM-based perceptual tokens	Integrated VLM	Diffusion transformer (DiT)	Continuous via diffusion
SeeDo	GPT-4o visual perception	GPT-4o with CoT prompting	VLM reasoning module	Text tokens
A^2	MaskCLIP	Vision-language priors	Cross-attention alignment	Continuous control signals
OpenVLA	DINOv2 + SigLIP fusion	Llama 2 backbone	Discretized action tokens	256-bin discretization
UP-VLA	CLIP-ViT + VQ-GAN	Phi-1.5	Policy head with MAP module	Low-level action output
RT-2	ViT-22B (PaLI-X)	PaLM-E decoder	Integrated with LLM	256-bin discretized text tokens
Generic VLA	ViT or DINOv2	PaLM or LLaMA	Diffusion or direct policy	End-effector velocities or joint torques

A fundamental architectural distinction emerges between models that directly repurpose VLMs for action prediction through simple quantization versus those employing specialized action modules. CogACT's componentized architecture with diffusion action transformers represents a departure from approaches that treat actions simply as discretized tokens . This design choice addresses the probabilistic and multimodal nature of robotic actions that regression-based schemes overlook . In contrast, RT-2 and OpenVLA represent actions as text tokens discretized into 256 bins, enabling direct integration into the language model framework .

The fused vision encoder approach combining DINOv2 and SigLIP features, as employed by OpenVLA, provides both semantic and spatial understanding capabilities . This architectural innovation addresses the limitation that VLMs often focus on high-level semantic content while neglecting low-level features crucial for detailed spatial information .

Training Methodologies

Training approaches vary substantially across VLA models, with implications for data efficiency and generalization. Table 3 summarizes training characteristics.

Model	Training Data Type	Data Scale	Training Procedure	Robot Platforms
CogACT	Real-world trajectories	>1M trajectories, 22.5M frames	Transfer learning with pre-trained modules	22 embodiments from Open X-Embodiment
SeeDo	Human demonstration videos	Not mentioned	Pipeline integration	UR10E robot arm
A^2	Simulated trajectories	~6.5k samples	Cross-entropy loss training	UR5 arm in PyBullet

Model	Training Data Type	Data Scale	Training Procedure	Robot Platforms
OpenVLA	Real-world demonstrations	970k trajectories	VLM backbone fine-tuning	Multiple from Open X-Embodiment
UP-VLA	Mixed robotic and image-text	25k robotic + 665k image-text	Two-stage co-fine-tuning	Franka-Emika Panda
RT-2	Web + robotic trajectories	Internet-scale + robotic data	Co-fine-tuning	13 robots in office kitchen
TPM	Real-world trajectories	1000 episodes	Behavior cloning	Franka Emika Research 3

A striking contrast exists between data requirements. A² achieves competitive performance with only ~6.5k samples through efficient alignment of action priors , while OpenVLA leverages 970k real-world trajectories and RT-2 combines Internet-scale data with robotic demonstrations . The vast scale of training data required for foundation models far exceeds what is feasible for many robot interactions , highlighting the importance of sample-efficient approaches.

Co-fine-tuning on both robotic trajectory data and Internet-scale vision-language tasks emerges as a key strategy for models like RT-2 and UP-VLA . This approach enables transfer of generalizable concepts from web data while maintaining task-specific robotic capabilities.

Pick-and-Place Task Characteristics

The reviewed studies evaluate VLA models across diverse pick-and-place scenarios with varying complexity. Table 4 characterizes the task settings.

Study	Object Types	Task Complexity	Environment	Language Instructions
CogACT	Everyday objects (fruits, tools)	Multi-step	Cluttered with distractors	Natural language
SeeDo	Vegetables, garments, blocks	Multi-step with temporal order	Structured containers	Video demonstration-based
A²	Novel objects, categories	Multi-step with obstacle clearing	Cluttered with occlusion	Labels, categories, colors, shapes
OpenVLA	Everyday + novel objects	Single to multi-step	Cluttered with distractors	Natural language
UP-VLA	Toy fruits, blocks	Single and multi-step	Cluttered, varied backgrounds	Natural language
RT-2	Everyday, novel, toys	Single and multi-step	Cluttered and clean	Natural language
TPM	Everyday objects	Single-step	Clean to cluttered	Natural language

Task complexity varies from simple single-step pick-and-place operations to long-horizon multi-step tasks requiring temporal reasoning and obstacle manipulation. All models struggle with complex manipulation tasks requiring

multi-step planning , representing a consistent challenge across the field. Environmental settings range from clean tabletop scenarios to cluttered configurations with multiple distractor objects and partial occlusions .

Performance Results

Success Rates and Comparative Performance

Quantitative performance metrics reveal substantial variation across models and evaluation conditions. Table 5 presents key performance results.

Model	Primary Metric	Performance	Baseline Comparison	Generalization Performance
CogACT	Success rate	74.8% (Visual Matching), 61.3% (Variant Aggregation)	+35% vs OpenVLA (sim), +55% (real)	Strong on unseen objects/backgrounds
SeeDo	Step Success Rate	>70% (daily tasks), >50% (block stacking)	Outperforms video VLM baselines	Strong zero-shot generalization
A^2	Task success rate	Higher than baselines with fewer steps	Outperforms neural field-based policies	Zero-shot to unseen objects/instructions
OpenVLA	Task success rate	+16.5% absolute vs RT-2-X	Outperforms RT-2-X and Octo	Strong across multiple embodiments
UP-VLA	Calvin ABC-D	+33% vs previous SOTA	Outperforms RT-1, Diffusion Policy	Better visual-semantic generalization
RT-2	Success rate	2x-6x improvement on generalization	3x baseline success rate	2x-3x better than RT-1 on new instructions
GPT-4o	NAMSE	Consistent low values (0.030-0.074)	Most consistent across tasks	Good zero-shot performance

CogACT demonstrates substantial improvements over OpenVLA, exceeding average success rates by over 35% in simulated evaluation and 55% in real robot experiments . It also outperforms the larger RT-2-X model (55B parameters) by 18% absolute success rates in simulation despite having approximately 7B parameters . OpenVLA similarly outperforms RT-2-X by 16.5% in absolute task success rate across 29 tasks with 7x fewer parameters .

UP-VLA achieves a 33% improvement on the Calvin ABC-D benchmark compared to previous state-of-the-art methods , particularly excelling in tasks requiring precise spatial information . RT-2 demonstrates 2x to 6x improvement over baselines in generalization tasks , with the PaLM-E version performing better in harder generalization scenarios while the PaLI-X version excels in easier ones .

Execution Efficiency

Computational efficiency varies across models. A² achieves inference times of approximately 1.0 seconds . RT-2's largest model runs at 1-3 Hz, while the smaller version operates at around 5 Hz . OpenVLA can be quantized to reduce memory footprint without compromising performance , though inference speed remains a concern for high-frequency control setups .

Generalization Capabilities

Generalization represents a critical evaluation dimension for VLA models. Table 6 synthesizes evidence across generalization categories.

Model	Unseen Objects	New Environments	Cross-Robot Transfer	Unseen Instructions
CogACT	Strong	Strong with camera repositioning	Tested on Realman and Franka	Implied
SeeDo	Some capability	Robust to environmental changes	Not mentioned	Not addressed
A²	Higher success on unseen	Novel camera viewpoints	Single camera real-world	Flexible language handling
OpenVLA	Demonstrated	Multiple platforms tested	WidowX and Google robot	Language grounding tasks
UP-VLA	Better visual-semantic	Diverse backgrounds	Not mentioned	Precise operation tasks
RT-2	Improved to novel	Outperforms baselines	Not explicitly mentioned	Commands not in training
Generic VLA	Designed for this	Some level of adaptation	RT-2: zero-shot transfer	Limited without fine-tuning

RT-2 demonstrates emergent capabilities including the ability to interpret commands not present in the robot training data, such as placing objects near semantically indicated locations (specific numbers or icons) and performing rudimentary reasoning (selecting smallest/largest objects) . Chain-of-thought reasoning enables RT-2 to perform multi-stage semantic reasoning, such as identifying appropriate improvised tools .

Cross-robot platform transfer capabilities vary considerably. RT-2 achieves zero-shot transfer across multiple robots and tasks , while OpenVLA demonstrates effectiveness across WidowX and Google robot platforms . However, zero-shot generalization to entirely novel robot models remains challenging and typically requires fine-tuning .

Evaluation Methodologies

The reviewed studies employ diverse evaluation setups. Table 7 summarizes experimental configurations.

Study	Robot Platform	Simulation Environment	Evaluation Trials	Camera Setup
CogACT	Google, WidowX, Realman, Franka	SIMPLER	24 (Pick), 11 (Franka tasks)	Intel RealSense
SeeDo	UR10E	Yes (unspecified)	Not specified	Intel RealSense 455
A^2	UR5 with ROBOTIQ-85	PyBullet	15 runs per test, 50 total	Intel RealSense L515
OpenVLA	WidowX, Google, Franka	LIBERO benchmark	170 (BridgeData), 60 (Google)	Third-person camera
UP-VLA	Franka-Emika Panda	Calvin ABC-D	>2k demonstrations	Visual observations
RT-2	7DoF mobile manipulator	Language-Table	6,000 trajectories	Vision-language inputs
Benchmark study	Open X-Embodiment robots	N/A (real data)	20 datasets	Images + language

RT-2's evaluation stands out for its scale with 6,000 evaluation trajectories , providing robust statistical assessment. The benchmarking study by Guruprasad et al. evaluates across 20 diverse datasets from the Open-X-Embodiment collection , enabling systematic comparison across varied manipulation tasks. Most studies employ RGB-D cameras (Intel RealSense variants) , with evaluation in both simulation and real-world settings.

Critical Factors Influencing Performance

Architectural Choices

The componentized VLA architecture with specialized action modules conditioned on VLM output significantly improves task performance and generalization compared to approaches that directly repurpose VLMs . The decoupling of cognition and action capabilities emerges as a key design principle . Diffusion action transformers for action sequence modeling demonstrate favorable scaling behaviors, with modest increases in parameters yielding significant performance gains .

The fused vision encoder combining DINOv2 and SigLIP improves spatial reasoning capabilities in OpenVLA . The combination of visual encoders provides robust visual features contributing to consistent performance . Using segmentation masks is more effective than bounding boxes for action prediction, and tracking is more robust than frame-by-frame detection for prompt generation .

Training Data and Procedures

Training data diversity and careful preprocessing are crucial for performance . The integration of both multi-modal understanding and future prediction objectives enhances high-level semantic comprehension and low-level spatial understanding simultaneously . Co-fine-tuning with mixed robotic trajectory and visual-language data enables transfer of generalizable concepts while maintaining task-specific capabilities .

Large-scale, diverse, and multi-modal datasets are emphasized as critical for effective VLA models . However, the vast scale of training data required for foundation models exceeds what is feasible for robot interactions , motivating

approaches like A^2 that achieve competitive performance with substantially less data through efficient alignment of action priors .

Prompt Engineering and Language Reasoning

Sophisticated prompt engineering emerges as a critical factor, with GPT-4o demonstrating the most consistent performance through this mechanism . The insertion of language reasoning structures to bridge high-level tasks to low-level actions enhances accuracy of action predictions . Chain-of-thought prompting helps generate task planning steps and improves spatial reasoning .

Identified Limitations and Failure Modes

Task-Specific Challenges

Complex manipulation tasks requiring multi-step planning or precise control remain challenging for all models . Tasks with strict place constraints in clutter present particular difficulties . Contact-rich skills require extensive demonstrations , and current models do not acquire new physical skills beyond those in training data .

Environmental and Perceptual Limitations

Heavy occlusion and visual ambiguity of target objects cause failures . Semantic ambiguity in language instructions leads to incorrect object selection . Spatial errors constitute a primary source of failure, attributed to limited spatial intelligence of current VLMs and imperfect tracking . Variable lighting, novel objects, or cluttered environments challenge generalization .

Computational and Data Constraints

High computational requirements characterize current approaches. OpenVLA fine-tuning requires 64 A100 GPUs for 14 days , and CogACT uses 16 NVIDIA A100 GPUs with 7.5 hours for fine-tuning . Integration of foundation models introduces latency issues affecting execution speed . OpenVLA's inference speed is problematic for high-frequency control setups .

Current VLMs focus on high-level semantic content but neglect low-level features crucial for detailed spatial information and physical dynamics . Models sometimes inaccurately identify specific objects due to data and backbone constraints . The gap between text output of VLMs and numerical predictions of robot actions remains a technical challenge .

Simulation-to-Real Transfer

Simulators simplify physical interactions, leading to poor transferability to real-world scenarios with complex dynamics . High-fidelity simulation platforms have low frame rates and high GPU demand, limiting their suitability for large-scale learning . The domain gap between real-world and simulated environments affects model performance when fine-tuned on simulated tasks .

Synthesis

The apparent heterogeneity in reported performance—with success rates ranging from approximately 50% to over 90% across studies—can be reconciled through examination of task complexity, evaluation conditions, and architectural choices.

Task Complexity and Environmental Factors

Studies reporting higher success rates typically evaluate on structured single-step tasks or within narrow task distributions. CogACT's 74.8% success rate in Visual Matching and SeeDo's >70% SSR for daily tasks occur in relatively controlled settings. Lower performance emerges consistently in cluttered environments with occlusions , multi-step planning requirements , and tasks involving precise spatial relationships . Block stacking tasks, which require precise spatial reasoning, show notably lower success rates (~50%) compared to simpler pick-and-place operations.

Architectural Design Trade-offs

The tension between model scale and efficiency reflects distinct design philosophies. RT-2's massive scale (tens of billions of parameters) enables emergent semantic reasoning capabilities but introduces computational constraints limiting deployment to 1-3 Hz inference . In contrast, OpenVLA and CogACT achieve competitive or superior performance with approximately 7B parameters through architectural innovations: componentized action modules and fused vision encoders . The 7B-parameter models outperform the 55B RT-2-X by 16-18% in absolute success rates , suggesting that specialized architectural design provides greater benefits than raw scale for manipulation tasks.

Models employing diffusion-based action prediction address the multimodal nature of robotic actions more effectively than simple discretization approaches , explaining their improved performance on tasks with multiple valid solutions. The action prior alignment approach of A^2 achieves competitive performance with only ~6.5k samples versus 970k for OpenVLA , demonstrating that efficient integration of foundation priors can dramatically reduce data requirements.

Generalization Boundaries

Cross-study comparison reveals consistent patterns in generalization limitations. All models demonstrate stronger performance on visual and object-level generalization (novel objects, backgrounds) compared to semantic or instruction-level generalization (novel commands, complex reasoning) . RT-2's emergent capabilities in interpreting unseen commands appear uniquely enabled by its massive-scale web pretraining, a resource unavailable to smaller models. Zero-shot cross-robot transfer remains achievable only with fine-tuning for most architectures , with RT-2 representing an exception through its foundation model heritage.

The consistent finding that spatial reasoning and precise manipulation remain challenging suggests that current vision encoders, optimized for semantic understanding, inadequately capture fine-grained spatial relationships. UP-VLA's integration of future visual prediction to enhance spatial understanding represents one approach to addressing this limitation, achieving 33% improvement on spatially demanding tasks .

References

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, K. Choromanski, Tianli Ding, Danny Driess, et al. “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control.” *Conference on Robot*

Learning, 2023.

Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. "VLM See, Robot Do: Human Demo Video to Robot Action Plan via Vision Language Model." *arXiv.org*, 2024.

Jiange Yang, Wenhui Tan, Chuhao Jin, Bei Liu, Jianlong Fu, Ruihua Song, and Limin Wang. "Pave the Way to Grasp Anything: Transferring Foundation Models for Universal Pick-Place Robots." *arXiv.org*, 2023.

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiangpei Zhu, and Jianyu Chen. "UP-VLA: A Unified Understanding and Prediction Model for Embodied Agent." *arXiv.org*, 2025.

Kechun Xu, Xunlong Xia, Kaixuan Wang, Yifei Yang, Yunxuan Mao, Bing Deng, Rong Xiong, and Yue Wang. "Efficient Alignment of Unconditioned Action Prior for Language-Conditioned Pick and Place in Clutter." *IEEE Transactions on Automation Science and Engineering*, 2025.

Lingling Fan, Kang Chen, Zhezhuang Xu, Meng Yuan, Ping Huang, and Weibing Huang. "Language Reasoning in Vision-Language-Action Model for Robotic Grasping." *ACM Cloud and Autonomic Computing Conference*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, A. Balakrishna, Suraj Nair, Rafael Rafailov, et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *Conference on Robot Learning*, 2024.

Muhayy ud Din, Waseem Akram, L. S. Saoud, Jan Rosell, and Irfan Hussain. "Vision Language Action Models in Robotic Manipulation: A Systematic Review." *arXiv.org*, 2025.

P. Guruprasad, Harshvardhan Digvijay Sikka, Jaewoo Song, Yangyue Wang, and Paul Pu Liang. "Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks." *arXiv.org*, 2024.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, M. Liao, Fangyun Wei, et al. "CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation." *arXiv.org*, 2024.