2025 17th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) | 979-8-3315-6759-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/IHMSC66529.2025.00046

Vision-Language-Action Models: Foundations, Techniques and Applications

Yan Li
School of Design and Fashion
Zhejiang University of Science and Technology
Hangzhou 310023 P.R. China
Email: liyan@zust.edu.cn

***Abstract*—Vision-Language-Action (VLA) models mark a transformative breakthrough in embodied AI, seamlessly integrating visual perception, natural language understanding, and robotic control to create intelligent, adaptable agents. Leveraging pre-trained vision-language models (VLMs), VLAs harness Internet-scale semantic knowledge for robust generalization across novel tasks, including object manipulation, navigation, and human interaction. This paper offers a survey of VLA models, elucidating their core concepts, architectural designs, and wide-ranging applications. We examine key architectural elements such as multimodal encoders, fusion mechanisms, and action decoders. Applications encompass robotic manipulation, autonomous navigation, human assistance, and emerging areas including edge computing, industrial automation, healthcare, agriculture, and virtual reality. By synthesizing these advancements, we highlight VLAs' potential to transcend traditional boundaries in robotics, while addressing persistent challenges in tokenization, fusion robustness, and embodiment generalization. This work provides a reference for researchers and practitioners, emphasizing VLAs' promise in developing scalable, versatile systems for real-world embodied intelligence.**

***Index Terms*—Vision-Language-Action Models, VLA architecture, VLA applications**

I. Introduction

In the rapidly evolving field of artificial intelligence and robotics, the quest for embodied agents capable of perceiving, reasoning, and acting in complex, dynamic environments has long been a central challenge. Traditional robotic systems often operate in silos: visual perception modules process sensory data in isolation, language understanding handles human commands separately, and control policies execute actions without deep semantic integration. This fragmentation limits generalization, adaptability, and efficiency, particularly in unstructured real-world scenarios where robots must interpret ambiguous instructions, navigate novel environments, and perform dexterous tasks [1], [2]. Vision-Language-Action (VLA) models emerge as a transformative paradigm to address these limitations, unifying visual perception, natural language comprehension, and embodied control within a cohesive learning framework [1], [3], [4].

At their core, VLA models build upon pre-trained vision-language models (VLMs), inheriting vast semantic knowledge from Internet-scale datasets to bridge the semantic gap between multimodal inputs (e.g., images and textual commands) and low-level robotic actions [2], [5]. This integration enables robots to not only perceive their surroundings but also understand high-level directives and generate precise, context-aware actions, fostering instruction-driven autonomy [6], [7]. For instance, VLAs facilitate rudimentary reasoning, such as selecting objects based on attributes or improvising tools in multi-stage tasks, while supporting closed-loop control in dynamic settings [4]. Unlike prior approaches reliant on hand-engineered features or isolated modules, VLAs offer superior versatility, dexterity, and generalizability, extending from controlled laboratories to everyday applications like home cleaning and industrial assembly [2], [3]. By transferring web-sourced knowledge to robotic control, these models empower agents to handle unseen objects, commands, and environments, marking a shift toward scalable, interpretable, and data-efficient systems [1], [8].

The motivation for VLA models stems from the growing demand for intelligent agents in diverse domains, where human-like multimodal reasoning is essential. In manipulation tasks, for example, robots must generalize to novel configurations; in navigation, they need to translate language goals into safe trajectories; and in human-robot interaction, they must ensure safety and responsiveness [6], [9]. However, challenges persist, including efficient multimodal fusion, robust tokenization for unified representations, generalization across robot embodiments, and generation of smooth, continuous motions [10], [11]. Addressing these requires a deep understanding of VLA concepts, architectures, and applications, which this paper aims to provide.

This paper is organized as follows. Section 2 introduces the foundational concepts of VLA models, outlining their objectives, capabilities, and integration of diverse modalities for real-world applicability. Section 3 delves into the architecture of VLA models, detailing multimodal input encoders, fusion mechanisms, action decoders, architectural paradigms, and tokenization strategies. Section 4 explores the broad applications of VLAs, spanning robotic manipulation, autonomous mobility, human assistance, specialized hardware, virtual environments, edge deployment, industrial robotics, healthcare, agriculture, and interactive navigation. Finally, Section 5 discusses open challenges and future research directions, emphasizing the potential of VLAs to revolutionize robotics.

2157-8982/25/$31.00 ©2025 IEEE
DOI 10.1109/IHMSC66529.2025.00046

II. Concepts of VLA

Vision-Language-Action (VLA) models mark a transformative paradigm in robotics, aiming to unify visual perception, natural language understanding, and embodied control within a single, cohesive learning framework [1], [2]. The core objectives and capabilities of VLA models encompass the following.

Integrating diverse modalities: VLAs seek to seamlessly fuse visual inputs with linguistic instructions to produce executable actions, effectively bridging perception, planning, and decision-making [6]. By combining visual perception, language comprehension, and motor control, these models empower embodied agents to interpret their environments, process commands, and execute precise actions. A pivotal goal is to transfer internet-scale semantic knowledge to robotic control, enabling superior generalization to novel objects and commands absent from training data [2]. This facilitates rudimentary reasoning (e.g., selecting the smallest object) and multi-stage semantic inference (e.g., identifying an improvised hammer). VLAs are engineered to map robot observations directly to actions, leveraging large-scale pre-training on web-sourced language and vision-language data, which supports closed-loop control in dynamic settings [4].

Enabling instruction-driven autonomy: VLAs empower robots to execute tasks via natural language prompts or directives from humans or high-level vision-language model (VLM) policies [1]. This includes "visual thinking" capabilities, where models generate subgoal images as intermediate reasoning steps to plan task execution [7]. Compared to prior methods, VLAs deliver greater versatility, dexterity, and generalizability in complex environments, spanning controlled labs to everyday households [5]. They prioritize high precision and multimodal integration, crucial for high-frequency dexterous operations [3].

Ultimately, VLAs target real-world applicability beyond laboratory confines, with broad generalization to unseen scenarios, such as navigation, manipulation, home cleaning, and multi-stage tasks [2], [8]. By overcoming the longstanding silos between visual recognition, language processing, and motor execution, VLAs offer a scalable, interpretable, and data-efficient alternative to traditional autonomous systems [2].

Typically constructed atop pre-trained vision-language models (VLMs), VLAs inherit expansive semantic knowledge and capitalize on established architectures for efficient training and deployment, while directly benefiting from ongoing VLM advancements [2].

III. Architecture of VLA

By building upon pre-trained vision-language models (VLMs), VLAs inherit vast semantic knowledge from Internet-scale datasets, enabling robots to perceive environments, interpret commands, and execute actions in a coherent manner [2]–[5], [10]. This holistic approach bridges the traditional divide between perception, cognition, and actuation, fostering more versatile and adaptable embodied agents [1]. At their core, VLA architectures consist of multimodal input encoders, fusion mechanisms, and action decoders, often underpinned by Transformer-based backbones. This section delineates the key components and design paradigms of VLA models, highlighting their innovations and variations.

The foundation of a VLA model lies in its multimodal input encoders, which process diverse data streams to create aligned representations. The visual encoder handles image-based inputs, such as RGB frames or depth maps, transforming them into fixed-length feature tokens. Popular choices include Transformer-based models like Vision Transformers (ViT), DINOv2, SigLIP, and CLIP, as well as convolutional networks like ResNet and EfficientNet. For instance, OpenVLA employs a fused encoder combining DINOv2 and SigLIP features [2], while π_0 leverages the PaliGemma VLM backbone [3], and DreamVLA uses a Masked Autoencoder for spatio-temporal patch representations [11]. Complementing this, the language encoder tokenizes and embeds natural language instructions—ranging from high-level goals to granular directives—into a shared feature space. Common implementations draw from large language models (LLMs) such as LLaMA, GPT, Qwen, or BERT [2], with OpenVLA specifically utilizing a LLaMA 2 7B backbone [9]. Additionally, the proprioceptive or state encoder captures the robot's internal state, including joint angles, end-effector poses, and gripper status, typically via multi-layer perceptrons (MLPs) or small Transformers. This enables reasoning about physical constraints, such as reachability and collision avoidance, while facilitating real-time feedback [1], [2].

A pivotal element of VLA architectures is multimodal fusion, which integrates visual, linguistic, and state information into a cohesive representation space [1], [2], [5], [8]. Techniques such as cross-modal attention, concatenated embeddings, or unified tokenization ensure seamless alignment. For example, π_0 adopts a late fusion strategy by embedding image observations into the language token space [3], whereas RT-2 tokenizes actions as "multimodal sentences" to adapt VLMs for direct robot control [8]. DreamVLA further refines this with structured causal and non-causal attention to disentangle various forms of world knowledge [11]. Such fusion mechanisms allow VLAs to synthesize multimodal inputs effectively, enabling context-aware decision-making.

Following fusion, the action decoder translates integrated embeddings into executable robot commands. This component varies widely to suit different control needs. Diffusion-based policies, employed in models like π_0, DiffusionVLA, TinyVLA, and CogACT, generate continuous, smooth actions using flow matching or diffusion processes, supporting high-frequency control (e.g., up to 50 Hz) and dexterous tasks [3], [8], [10]–[12]. DreamVLA, for instance, incorporates a Denoising Diffusion Transformer [11]. Alternatively, autoregressive Transformer heads, as in RT-2, output low-level actions as tokenized text [4], while OpenVLA predicts discretized action tokens via a LLaMA 2 head [1]. CoT-VLA extends this to action chunking for sequential predictions [7]. Simpler variants, such as MLP or latent-conditioned policy heads, appear in models like iRe-VLA and HiRT [1], prioritizing

efficiency over complexity.

VLA models exhibit diverse architectural paradigms, reflecting trade-offs in design philosophy. End-to-end approaches, such as CLIPort, RT-1, OpenVLA, and ShowUI-2B, process raw inputs directly into motor commands through a single network, promoting seamless integration. In contrast, component-focused designs, like VLATest and Chain-of-Affordance, modularize perception, language grounding, and action, allowing targeted optimizations. Hierarchical structures address long-horizon tasks by decoupling high-level planning from low-level execution; for example, CogACT and NaVILA use LLM-based planners to generate subgoals for reactive controllers, and $\pi_{0.5}$ infers subtasks before predicting actions [8]. Flat policies, conversely, map inputs to actions without intermediate layers. Other paradigms include early fusion (e.g., EF-VLA) for preserving pre-training alignments, dual-system setups (e.g., NVIDIA's GR00T N1) combining reactive and deliberative control, and self-correcting frameworks (e.g., SC-VLA) for autonomous error recovery. The Transformer architecture underpins most VLAs, providing scalability and expressiveness.

Central to VLA innovation is tokenization and world encoding, which unifies modalities into a shared embedding space. Prefix tokens encode scenes (e.g., images or videos) and instructions, establishing goal and environmental context. State tokens integrate real-time proprioception, fused with prefixes to reason about constraints. Action tokens are then autoregressively generated to produce motor sequences, as exemplified by RT-2 and PaLM-E [4]. This token-based framework facilitates end-to-end training and generalization.

VLA architectures bridge visual, linguistic, and action domains, advancing robotic dexterity and adaptability. Nonetheless, challenges persist in efficient tokenization, robust fusion, embodiment generalization, and continuous motion generation, warranting ongoing research.

IV. Applications of VLA

Vision-Language-Action (VLA) models mark a significant advancement in embodied AI, enabling robots and agents to seamlessly integrate environmental perception, natural language interpretation, and precise action execution. By bridging the semantic divide between multimodal inputs—such as images and textual commands—and low-level robotic control, VLAs facilitate generalization across diverse tasks, including object manipulation, navigation, and human interaction. This unified approach empowers deployment in real-world scenarios ranging from routine household chores to specialized industrial and medical operations, fostering adaptable and intelligent systems that respond dynamically to complex environments.

A primary domain for VLA applications is robotic manipulation and task generalization, where models handle object-level interactions from basic grasping to intricate assemblies, often adapting to novel objects and settings. For instance, π_0 excels in dexterous and long-horizon tasks, such as laundry folding, table cleaning, and full-room organization like kitchen or bedroom tidying [3]. OpenVLA demonstrates superior performance in multi-object manipulation and language-grounded tasks, including pick-and-place operations (e.g., "Put Carrot in Bowl" or "Move object onto Plate") on platforms like Franka robots [9]. RT-2 leverages web-scale knowledge for reasoning-intensive tasks, such as relocating objects based on color or symbolic cues [4]. Similarly, TinyVLA supports multi-task and bimanual manipulation, executing actions like mug flipping and cube stacking [10]. Other models enhance specificity: Chain-of-Affordance (CoA) aids obstacle avoidance and free-space identification [7]; TLA integrates tactile sensing for contact-rich tasks like peg-in-hole insertion [1]; ChatVLA manages 25 diverse real-world tasks across environments like kitchens and tabletops [13]; and DreamVLA improves action reasoning through world knowledge forecasting [11]. These capabilities underscore VLAs' role in scalable, generalist manipulation.

In autonomous mobility, VLAs translate high-level language goals into safe navigation and control strategies for various robotic platforms, including wheeled, legged, and aerial systems. QUAR-VLA, for example, fuses visual data and instructions to enable quadruped robots to perform goal-oriented navigation and whole-body manipulation. For vehicular applications, CoVLA aligns perceptual and linguistic features with driving actions, while OpenDriveVLA provides end-to-end trajectory planning in urban settings using multi-view vision and language inputs. UAV-VLA supports zero-shot aerial mission planning by combining satellite imagery with natural language directives. Additionally, ADAPT enhances vision-language navigation (VLN) by refining action-level modality alignments, aiding agents in complex visual scenarios. These models highlight VLAs' efficacy in dynamic, mobility-centric tasks.

Human assistance and interaction represent another critical application area, where VLAs interpret commands and contexts to facilitate collaborative, safe engagements. RoboNurse-VLA offers a real-time voice-to-action system for surgical instrument handover, robustly managing unseen tools in dynamic operating rooms. ChatVLA integrates conversational skills with multimodal understanding and control, performing well in visual question answering and interactive tasks [13]. Extending to specialized hardware, models like Helix[1] enable high-degree-of-freedom humanoid control for tasks such as item retrieval from refrigerators, supporting zero-shot transfers. GR00T N1, a diffusion-based model, unifies control for humanoid platforms. In virtual domains, VLAs automate digital interfaces: ShowUI [14] handles GUI tasks, and JARVIS-VLA [15] predicts inputs for open-world 3D games.

Efficiency-focused applications adapt VLAs for edge and low-power scenarios, ensuring real-time performance on constrained hardware. Edge VLA achieves high inference speeds on edge devices, while SmolVLA enables CPU-based real-time operation. TinyVLA further optimizes for data effi-

[1] https://www.figure.ai/news/helix

ciency and speed without heavy pre-training [10]. In industrial robotics, CogACT supports precise assembly and adaptation in manufacturing [16]. Healthcare applications leverage VLAs for high-stakes tasks, fusing video feeds, anatomical data, and commands for procedures like sub-millimeter suturing. In agriculture, VLA-equipped systems perform vision-guided harvesting, identifying ripe produce based on criteria like "pick only Grade A fruits" and executing precise actions or drone-guided irrigation. Interactive augmented reality (AR) navigation also benefits, with VLAs processing queries (e.g., "avoid stairs to Gate 22") and visual cues to adapt paths in environments like airports.

VLA models unify disparate perception and action components into semantically aligned, instruction-following agents, driving innovation across manipulation, mobility, human collaboration, and beyond. Their adaptability enhances robustness in diverse scenarios, from industrial precision to everyday assistance, while ongoing developments promise even broader real-world impact.

V. Discussion and Conclusion

The advent of Vision-Language-Action (VLA) models marks a pivotal evolution in embodied AI, integrating multimodal perception, linguistic reasoning, and action execution to surpass traditional robotic paradigms with enhanced adaptability and intelligence. From foundational concepts emphasizing modality fusion and autonomy, to innovative architectures like Transformer encoders and action decoders, VLAs enable applications in manipulation, navigation, human interaction, and specialized fields like healthcare and agriculture. This approach leverages pre-trained VLMs for semantic generalization, outperforming isolated systems in versatility and efficiency. However, limitations persist, including heavy reliance on resource-intensive pre-training, performance degradation in noisy or adversarial environments, and ethical concerns like bias inheritance. Open challenges include efficient tokenization, robust multimodal fusion, embodiment generalization, and smooth motion generation, often leading to latency, imbalances, or instability in real-time or cross-platform scenarios. Future research should focus on enhancing data efficiency via few-shot learning, incorporating additional modalities like tactile inputs, and developing hierarchical self-correcting systems for long-horizon tasks. Edge-optimized models and interdisciplinary efforts in ethics and benchmarks will democratize VLAs for resource-constrained environments.

In conclusion, this paper provides an overview of VLA models, highlighting their transformative impact on unifying perception, cognition, and action for versatile agents. While challenges in tokenization, fusion, and generalization remain, ongoing advancements promise scalable robotic systems that redefine autonomous intelligence and foster seamless human-robot collaboration across domains.

Acknowledgement

This work is sponsored by Ningbo "Yongjiang Science Innovation 2035" Key Technology Breakthrough Plan (No. 2025Z056).

References

[1] M. U. Din, W. Akram, L. S. Saoud, J. Rosell, and I. Hussain, "Vision language action models in robotic manipulation: A systematic review," 2025.

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

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024.

[4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 2165–2183.

[5] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," 2024.

[6] P. Ding, H. Zhao, W. Zhang, W. Song, M. Zhang, S. Huang, N. Yang, and D. Wang, "Quar-vla: Vision-language-action model for quadruped robots," 2023.

[7] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, T.-Y. Lin, G. Wetzstein, M.-Y. Liu, and D. Xiang, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 1702–1713.

[8] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025.

[9] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," 2024.

[10] J. Wen, Y. Zhu, J. Li, M. Zhu, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng, F. Feng, and J. Tang, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," 2024.

[11] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, H. Wang, Z. Zhang, L. Yi, W. Zeng, and X. Jin, "Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge," 2025.

[12] J. Li, Y. Zhu, Z. Tang, J. Wen, M. Zhu, X. Liu, C. Li, R. Cheng, Y. Peng, Y. Peng, and F. Feng, "Coa-vla: Improving vision-language-action models via visual-textual chain-of-affordance," 2024.

[13] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, R. Cheng, Y. Peng, C. Shen, and F. Feng, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," 2025.

[14] K. Q. Lin, L. Li, D. Gao, Z. Yang, S. Wu, Z. Bai, S. W. Lei, L. Wang, and M. Z. Shou, "Showui: One vision-language-action model for gui visual agent," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 19 498–19 508.

[15] M. Li, Z. Wang, K. He, X. Ma, and Y. Liang, "Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse," 2025.

[16] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, X. Wang, B. Liu, J. Fu, J. Bao, D. Chen, Y. Shi, J. Yang, and B. Guo, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," 2024.