From Perception to Action with Integrated VLA Systems

Rashid Turgunbaev
Kokand State University

Abstract: *The historical trajectory of robotics has been characterized by a profound compartmentalization of its constituent subsystems. Perception, world modeling, planning, and low-level control have evolved as largely independent fields, each with its own paradigms, benchmarks, and failure modes. The integration of these components into a cohesive, robust, and generalizable robotic agent has remained a persistent "grand challenge." The recent advent of large-scale Vision-Language-Action (VLA) models presents a paradigmatic shift, promising a move from this fragmented architecture toward a unified, end-to-end approach. This article examines the foundational principles, architectural innovations, and practical implications of integrated VLA systems that directly translate perceptual inputs and linguistic instructions into executable robotic actions. We argue that these systems do not merely incrementally improve individual components but fundamentally redefine the robot's cognitive pipeline, enabling a new level of semantic understanding, contextual adaptation, and open-world generalization. By treating the problem of embodied agency as a sequence modeling task across multimodal tokens, VLA models offer a path toward robots that can seamlessly connect the what they see and are told with the how they must physically act.*
Keywords: *embodied artificial intelligence, multimodal robot learning, vision-language-action models, end-to-end robotic control, semantic task execution, transformer-based robotics*

Introduction

For decades, the canonical robot control stack has been a sequential, modular cascade. A perception module, perhaps using classical computer vision or later convolutional neural networks, would process sensor data to produce geometric or semantic maps - a list of objects, their poses, and perhaps affordances. A separate planning system, leveraging symbolic AI or sampling-based algorithms, would use this abstracted world model to formulate a sequence of goals or waypoints to satisfy a command. Finally, a control layer, often grounded in dynamics and classical feedback loops, would translate these intermediate representations into motor torques or joint velocities. This pipeline, while logically sound, accumulates errors and suffers from brittle interfaces. Each module operates on lossy abstractions; the planner does not see the raw visual uncertainty, and the controller is blind to the semantic intent of the task. The system's performance is bottlenecked by the least reliable component and is notoriously difficult to adapt to novel objects, environments, or instructions. Generalization requires painstaking re-engineering at each level.

The rise of foundation models, particularly large language models (LLMs) and vision-language models (VLMs), has injected a transformative idea into this landscape: the power of scale and cross-modal pre-training. LLMs demonstrated that internet-scale text corpora could yield models with profound semantic knowledge and reasoning capabilities, while VLMs showed that aligning visual embeddings with linguistic ones could ground this knowledge in perceptual reality. The logical, and ambitious, extension is to close the loop by embodying these models - connecting their semantic and perceptual understanding to the physical dimension of action. This is the core proposition of Vision-Language-Action models. An integrated VLA system is not a loose coupling of an off-the-shelf VLM with a traditional robot planner. Instead, it is an architecture trained, either end-to-end or through careful alignment, to consume direct sensory observations (like images or proprioception) and textual commands, and to output low-level action sequences (like joint displacements or end-effector velocities) or actionable policy representations. This integration collapses the traditional stack, promising to mitigate abstraction loss and create a more fluid, context-aware, and trainable agent.

At the heart of the integrated VLA approach is a conceptual reframing of robotics as a multimodal sequence modeling problem. This paradigm draws direct inspiration from the success of transformer architectures in language and vision. In this view, a robot's experience is a sequential stream of interleaved tokens. A segment of this stream might consist of: an instruction token sequence ("pick up the blue block"), followed by a sequence of visual tokens representing the current image observation from the robot's camera, followed by a sequence of action tokens representing the robot's chosen motor command, followed by the next visual tokens, and so on.

The model's training objective is, fundamentally, prediction. Given a context of past instructions, observations, and actions, the model is trained to predict the next token in the sequence. When the next token is part of an action command, the model is effectively learning a control policy. Crucially, because it must also predict future visual tokens or interpret instruction tokens in context, it implicitly builds a world model - a representation that encodes not just geometry but physics and causality. It learns that certain action sequences lead to predictable changes in the visual scene. This is a form of self-supervised learning on a grand scale, where the data can consist of vast teleoperation datasets, simulation rollouts, or video archives of robotic manipulation.

Several architectural implementations exist. One prominent method is to treat actions as just another language. Here, a pre-trained LLM or VLM backbone is adapted by extending its vocabulary to include discrete "action tokens." These tokens can represent primitive motor commands or higher-level skills. During training on robot data, the model learns to "speak" in this action language, grounding its predictions in the visual and linguistic context. Another approach treats actions as a continuous modality. The visual and textual inputs are encoded into a shared embedding space by a transformer encoder, and a separate transformer decoder or a diffusion model then generates a sequence of continuous action vectors (e.g., delta poses for a gripper). Regardless of the specific instantiation, the core principle remains: a single, unified model processes all modalities within a shared representational space and generates the action output autoregressively or iteratively.

The power of an integrated VLA system lies in the emergent properties that arise from this tight coupling of perception, language, and action. Three advantages are particularly salient: dense semantic grounding, implicit affordance learning, and natural contextual adaptation.

First, semantic grounding moves from being a separate, intermediate task to a foundational property of the model's representations. In a traditional pipeline, an object detector might label a region as "mug." The planner must then access a separate knowledge base to understand that a mug is "graspable," "pour-able," and "fragile." In a VLA model, the visual representation of the mug is inherently linked, through its pre-training on internet-scale image-text data and its fine-tuning on action data, to all these semantic and functional properties. When the model processes the pixels of the mug in the context of the command "pour water," the activation patterns that correspond to linguistic concepts like "handle," "full," "tip," and "liquid" are inherently primed, directly influencing the action trajectory it generates. The perception is not neutral; it is already infused with actionable meaning.

Second, this leads to the implicit learning of affordances - the possibilities for action an environment offers. A VLA model does not need an explicit list of object affordances. By training on successful interaction data, it learns that pixel patterns corresponding to "chair" are predictive of successful "sit" action sequences, and those corresponding to "door handle" are predictive of successful "pull" or "turn" actions. The affordance is encoded in the correlation between the visual representation and the action output distribution. This allows for remarkable generalization to novel objects. A robot trained primarily on mugs and cups might successfully manipulate a novel, oddly-shaped vase if its visual features are deemed "sufficiently mug-like" by the model's representation space, because the functional embedding of "graspable container" is shared.

Third, integrated VLAs exhibit a form of contextual plasticity that is hard to engineer into modular systems. The meaning of an action is contingent on the full context. The motor command for a "gentle" push versus a "forceful" push is not a fixed parameter but is modulated by the linguistic modifier. The interpretation of "pick it up" depends entirely on what "it" is co-referenced to in the visual scene and dialogue history. In a VLA, the language instruction directly conditions the entire generative process, softly guiding the feature activations from the perception stream and biasing the action decoder. This enables nuanced, compositional instruction following. Commands like "put the apple next to the plate, but not too close to the knife" can be processed in one pass, as the model's linguistic understanding modulates spatial reasoning directly in the action-generating pathway, without needing a symbolic spatial relation parser.

Despite their transformative potential, the development and deployment of integrated VLA systems face significant hurdles. These challenges define the current frontiers of research.

A primary concern is the data scarcity and distribution shift problem. While VLMs and LLMs train on billions of web-scale image-text pairs and tokens, high-quality, diverse robot action data is orders of magnitude scarcer. Collecting physical robot data is slow, expensive, and lacks the breadth of the internet. This creates a mismatch between the model's rich semantic and visual knowledge and its limited understanding of physical interaction. Techniques to bridge this gap are critical, including: large-scale simulation with realistic rendering and physics to pre-train "visual motor controllers"; efficient fine-tuning methods like adapter networks or low-rank adaptation (LoRA) that lock most of the pre-trained knowledge and only train small action-specific modules; and novel data augmentation strategies that leverage internet videos to learn plausible action effects, even if the action itself is not performed.

Relatedly, the issue of safety and reliability is magnified. A monolithic neural network controlling a physical robot is a "black box" in the most consequential sense. Its failure modes can be unpredictable and catastrophic. Ensuring that such a system operates within safe physical constraints, respects human intentions, and can be audited or debugged is paramount. Research is needed into "safe-by-design" architectures, perhaps incorporating verifiable control barriers or allowing for natural language specification of constraints. The model must also be able to express uncertainty and know when a command is ambiguous or its own capabilities are insufficient, requiring a move from pure generative models to ones that can also query for help.

Finally, there is the challenge of temporal scale and compositionality. Current VLAs excel at short-horizon tasks (e.g., "pick and place") but struggle with long-horizon, multi-step problems that require intricate sequencing and state tracking over minutes or hours. This is partly a limitation of the sequence length transformers can handle effectively. Solving this may require hierarchical approaches, where a high-level VLA planner proposes sub-goals in language or a sketch of actions, which are then executed by a lower-level, integrated VLA controller. This creates a recursive integration, maintaining the semantic grounding across different temporal abstractions.

Conclusion

The pursuit of integrated Vision-Language-Action systems represents more than just a new tool in the roboticist's kit. It signifies a philosophical shift toward viewing embodied intelligence as an inseparable whole. The traditional modular pipeline, for all its engineering merits, imposes an artificial dichotomy between understanding and doing. VLA models, by learning to predict action tokens from perceptual and linguistic context, strive to unify these facets.

The path forward is not the abandonment of all classical robotics. Rather, the future likely lies in a synergistic fusion. The robust, certifiable principles of dynamics and control theory can provide a safety-critical substrate upon which VLA policies operate, much like the human cerebellum's regulation of the cerebral cortex's plans. The role of symbolic reasoning may evolve from being the central planner to being a high-level "prompt engineer" or verifier for the VLA's outputs.

"From Perception to Action" is thus both a technical description and a manifesto. Integrated VLA systems promise to create robots that do not just process visual data and execute parsed commands, but that understand scenes in terms of possibilities and act in accordance with rich, contextual meaning. By closing the loop between the high-level semantic world of language and the low-level physical world of forces and motions, these models offer our most promising yet challenging route toward robots that can flexibly, intelligently, and safely operate in the open-ended complexity of human environments. The journey from perception to action, long a fragmented and brittle chain, is being reforged into a continuous, learning, and adaptable circuit.

References

1. Ktaybekova, D. J. (2025). METHODOLOGICAL APPROACHES APPROPRIATE TO AGE CHARACTERISTICS IN THE ACQUISITION OF DUTOR CREATIVITY AND PERFORMANCE SKILLS. European Review of Contemporary Arts and Humanities, 1(5), 9-14.
2. Jurayev, I. (2025). A SYSTEMATIC PEDAGOGICAL APPROACH TO MAQOM ENCULTURATION IN UNDERGRADUATE VOCAL STUDIES. European Review of Contemporary Arts and Humanities, 1(5), 34-41.
3. Turgunbaev, R. (2025). CORE DATABASE COMPETENCIES FOR LIBRARY STUDENTS. European Review of Contemporary Arts and Humanities, 1(5), 104-111.
4. Berdikhanova, D. M. (2025). PEDAGOGICAL MECHANISMS FOR DEVELOPING STUDENTS'ARTISTIC THINKING IN DUTOR PERFORMANCE. European Review of Contemporary Arts and Humanities, 1(5), 4-8.
5. qizi Oxunjonova, G. Z. O. (2025). METHODS OF DEVELOPING STUDENTS'SPATIAL THINKING ABILITIES IN DRAWING LESSONS. European Review of Contemporary Arts and Humanities, 1(5), 87-91.
6. Xakimxo'ja qizi Odilova, M. (2025). MODERN APPROACH AND IMPORTANCE OF DRAWING TEACHING. European Review of Contemporary Arts and Humanities, 1(5), 83-86.
7. Sattarov, F. I. (2025). THE FORMATION OF COLORS IN NATURE AND THEIR CHARACTERISTICS OF CHANGE. European Review of Contemporary Arts and Humanities, 1(5), 70-73.
8. qizi Rahimjonova, N. B. (2025). EFFECTIVENESS OF USING DIGITAL TECHNOLOGIES IN DRAWING. European Review of Contemporary Arts and Humanities, 1(5), 78-82.
9. qizi Boyigitova, G. B. (2025). CONNECTIONS AND ITS TYPES. European Review of Contemporary Arts and Humanities, 1(5), 65-69.
10. qizi Xabibullayeva, K. Z. (2025). SHEARING AND ITS IMPORTANCE IN DRAWINGS. European Review of Contemporary Arts and Humanities, 1(5), 116-121.
11. Turg'unboy qizi Karimjonova, D. (2025). FINE ART OF UZBEKISTAN. European Review of Contemporary Arts and Humanities, 1(5), 29-33.
12. qizi Amoniddinova, I. B. (2025). RESTORATION OF PAINTING. European Review of Contemporary Arts and Humanities, 1(5), 51-54.
13. qizi G'aybullayeva, T. O. (2025). STAGES OF CREATING A COMPLEX STILL LIFE COMPOSITION IN PAINTING. European Review of Contemporary Arts and Humanities, 1(5), 74-77.
14. Kurbanova, M. M. (2025). COMPARATIVE ANALYSIS OF EASTERN AND WESTERN ORCHESTRAL TRADITIONS. European Review of Contemporary Arts and Humanities, 1(5), 55-59.
15. Isaqov, S. S. (2025). NONVERBAL COMMUNICATION AND COHESION IN THE ORCHESTRAL ENSEMBLE. European Review of Contemporary Arts and Humanities, 1(5), 112-115.

16. Ibragimov, S. (2025). TEACHING ACADEMIC WRITING WITH CHATGPT: OPPORTUNITIES AND LIMITATIONS. European Review of Contemporary Arts and Humanities, 1(5), 92-96.
17. Narmatova, A. P. (2025). DEVELOPING COUNTRY-SPECIFIC CULTURAL COMPETENCE IN GERMAN LANGUAGE LEARNERS THROUGH LITERARY TEXTS. European Review of Contemporary Arts and Humanities, 1(5), 97-103.
18. Qurbonova, S. (2025). ANALYSIS OF ORNAMENTATION IN UZBEK MAQOM VOCAL PERFORMANCE. European Review of Contemporary Arts and Humanities, 1(5), 15-19.
19. Muminov, M. (2025). TRANSMISSION OF SHASHMAQOM REPERTOIRE IN CONTEMPORARY MASTER-STUDENT RELATIONSHIPS. European Review of Contemporary Arts and Humanities, 1(5), 60-64.
20. Akhrorova, M. (2025). THE CONCEPT OF TRAGIC FATE IN COMPARATIVE LITERATURE: EASTERN AND WESTERN PERSPECTIVES. European Review of Contemporary Arts and Humanities, 1(5), 46-50.
21. Boboyev, V. (2025). MICROTONAL INTONATION AND ORNAMENTATION IN THE KASHKAR RUBAB MAQOM REPERTOIRE. European Review of Contemporary Arts and Humanities, 1(4), 35-44.
22. Mustafoev, S. M. (2025). THE INTERCONNECTION BETWEEN SOUNDS, MUSICAL MEMORY, AND THE SENSE OF MELODY AND HARMONY. European Review of Contemporary Arts and Humanities, 1(4), 3-8.
23. Egamberganova, Z. (2025). INTEGRATING RFID WITH SMART SHELVES AND ROBOTIC RETRIEVAL SYSTEMS FOR THE AUTONOMOUS LIBRARY. European Review of Contemporary Arts and Humanities, 1(4), 45-50.
24. Mustafoev, S. M. (2025). THE LOCALIZATION OF THE FRONTAL AND PARIETAL AREAS OF REPRODUCTION IN CLASSICAL ARTISTS AND MUSICIANS. European Review of Contemporary Arts and Humanities, 1(4), 9-13.
25. Ma’murjon qizi Khatamkulova, Z. (2025). CHALLENGES OF IMPLEMENTING STEAM IN ENGLISH LANGUAGE CLASSES. European Review of Contemporary Arts and Humanities, 1(4), 14-17.
26. Xo’jjiyevev, M. Y., & Bozorova, F. J. R. (2025). METROLOGICAL LIMITS OF ACCURACY OF PUMPKIN SEED OIL ADDITION TO FUNCTIONAL DRINKS. European Review of Contemporary Arts and Humanities, 1(4), 23-26.
27. ogli Oktyabrov, M. A. (2025). THE EMOTIONAL EXPRESSION OF ARTISTS THROUGH COLORS AND THE PSYCHOLOGICAL EFFECT OF COLORS IN ARTWORKS. European Review of Contemporary Arts and Humanities, 1(4), 30-34.
28. Abdunabiyeva, M. (2025). THE CULTURAL IDENTITY AND AESTHETIC EXPRESSION IN UZBEK NATIONAL DANCE ART. European Review of Contemporary Arts and Humanities, 1(3), 18-24.
29. Mirzaitova, M., & Astanakulov, O. (2025). CURRENT STATE OF INVESTMENT ACTIVITY IN TOURISM ORGANIZATIONS. European Review of Contemporary Arts and Humanities, 1(3), 14-17.
30. ogli Muqimov, S. Z. (2025). MUSIC AND NEUROPHYSIOLOGY: HOW DOES MUSIC CHANGE BRAIN ACTIVITY?. European Review of Contemporary Arts and Humanities, 1(3), 3-7.
31. oglu Muqimov, S. Z. (2025). INTERPRETING REPETITION AND VARIATION IN DIGITAL MUSIC: FROM ALGORITHMS TO ARTISTIC EXPRESSION. European Review of Contemporary Arts and Humanities, 1(3), 8-13.

32. Egamberdiyeva, Z. (2025). LIBRARIES AS CENTERS OF LIFELONG LEARNING AND COMMUNITY ENGAGEMENT. European Review of Contemporary Arts and Humanities, 1(2), 3-7.
33. Sharobiddinova, M. (2025). THE ROLE OF UZBEK MUSICAL INSTRUMENTS IN PEDAGOGY, PERFORMANCE, AND CULTURAL IDENTITY. European Review of Contemporary Arts and Humanities, 1(2), 12-16.
34. Turanov, D. A. (2025). PERSPECTIVES AND RISKS OF ARTIFICIAL INTELLIGENCE IN THE JUDICIAL SYSTEM OF UZBEKISTAN IN THE CONTEXT OF INTERNATIONAL EXPERIENCE. European Review of Contemporary Arts and Humanities, 1(2), 8-11.
35. Madaminjonovna, M., & Uzoqjonov, M. (2025). The possibilities of solving environmental problems through technology. Academic Journal of Science, Technology and Education, 1(7), 11-15.
36. Ergasheva, N. N., & kizi Ruzikulova, K. N. (2025). Features of the course of iron deficiency conditions and risk factors for their development in children of the first year of life. Academic Journal of Science, Technology and Education, 1(7), 26-29.
37. Shomuradova, D. A. (2025). Effective use of authentic materials in developing professional communicative skills. Academic Journal of Science, Technology and Education, 1(7), 53-55.
38. Ahmedov, A. B. (2025). The emergence of digital platforms in Uzbekistan’s economy and their impact on market competition. Academic Journal of Science, Technology and Education, 1(7), 6-10.
39. Ergasheva, N. N., & Ergashaliyev, S. D. (2025). The role of gut microflora in postoperative adaptation and necrotizing enterocolitis risk following congenital intestinal obstruction. Academic Journal of Science, Technology and Education, 1(7), 56-59.
40. Abdusalimova, A. (2025). Improving the effectiveness of fine arts lessons through innovative pedagogical technologies. Academic Journal of Science, Technology and Education, 1(7), 36-39.
41. Asatullayev, R. B., & Latifova, S. N. (2025). The lymphatic system. Academic Journal of Science, Technology and Education, 1(7), 43-44.
42. Jurayeva, M. A. (2025). The formation of Uzbek journalism at the beginning of the 20th century. Academic Journal of Science, Technology and Education, 1(7), 63-65.
43. Xo’jjiyevev, M. Y., & Bozorova, F. J. R. (2025). Antioxidant properties of pumpkin seed oil addition to functional drinks. Academic Journal of Science, Technology and Education, 1(7), 40-42.
44. Yuldashev, K. (2025). Genre and performance style in theater art. Academic Journal of Science, Technology and Education, 1(7), 45-49.
45. oglu Nematov, F. N. (2025). Building effective collaboration between teachers and parents in primary school. Academic Journal of Science, Technology and Education, 1(7), 60-62.
46. Turgunbaev, R. (2025). Enhancing student understanding of artificial intelligence through practical neural network applications. Academic Journal of Science, Technology and Education, 1(6), 36-40.
47. Arzikulov, F., & Komiljonov, A. (2025). The role of artificial intelligence in personalized oncology: predictive models and treatment optimization. Academic Journal of Science, Technology and Education, 1(6), 24-33.
48. Arzikulov, F., & Komiljonov, A. (2025). AI-powered diagnostic systems in radiology: enhancing precision, speed, and clinical decision-making. Academic Journal of Science, Technology and Education, 1(6), 16-23.
49. Abdusattarova, M. (2025). The art of leadership. Academic Journal of Science, Technology and Education, 1(6), 4-7.
50. Nasirova, G. R. (2025). Understanding structural changes in the respiratory tract in chronic disease. Academic Journal of Science, Technology and Education, 1(6), 82-85.

51. Masharipova, S. A. (2025). Prospects for the use of digitalization in the effective financial management of the chemical industry. Academic Journal of Science, Technology and Education, 1(6), 95-98.
52. Ergasheva, G. N. (2025). Storytelling method in teaching English to preschool children. Academic Journal of Science, Technology and Education, 1(6), 63-65.
53. Kamalova, A. (2025). Shukur Khollmirzaev's essay "That person is a mentor, and I am a disciple", its composition, plot, and system of characters. Academic Journal of Science, Technology and Education, 1(6), 48-50.
54. Mahmudova, D. Q. (2025). Effectiveness of microfertilizers in corn cultivation. Academic Journal of Science, Technology and Education, 1(6), 51-52.
55. Tokhtayev, I., & Ganiyeva, F. (2025). Opportunities for organic potato cultivation. Academic Journal of Science, Technology and Education, 1(6), 66-69.
56. Eraliyeva, M. T. K. (2025). Constitutional rights of citizens in the field of social security and their guarantees. Academic Journal of Science, Technology and Education, 1(6), 53-54.
57. oglu Idiyev, B. B., & Khujakulov, K. R. (2025). Synthesis and kinetic regularities of copolymers based on styrene and nitrogen-containing methacrylic monomers. Academic Journal of Science, Technology and Education, 1(6), 99-104.
58. qizi Haydarova, S. A. (2025). Electromagnetism. Academic Journal of Science, Technology and Education, 1(6), 34-35.
59. qizi Kenjayeva, Z. S. (2025). Advantages of modern methodology in forming phonetic competence in primary school students. Academic Journal of Science, Technology and Education, 1(6), 59-62.
60. qizi Xurramova, S. Q. (2025). Neurolinguistics: Comparative study of language processing in English and Uzbek. Academic Journal of Science, Technology and Education, 1(6), 12-15.
61. qizi Vakilova, S. T. (2025). Technological factors influencing the antioxidant activity of mulberry leaf tea. Academic Journal of Science, Technology and Education, 1(6), 45-47.
62. Egamberdiyeva, Z. (2025). Innovative approaches to teaching librarianship in the digital age. Academic Journal of Science, Technology and Education, 1(5), 7-11.
63. Ibatova, N. I. (2025). Pedagogical Approaches and Methods to Enhance the Effectiveness of Developing Students' Artistic and Figurative Thinking. Academic Journal of Science, Technology and Education, 1(5), 46-49.
64. Santos, A. J. (2025). The Role of Behavioral Insights in Economic Decision Making Education. Academic Journal of Science, Technology and Education, 1(5), 39-42.
65. oglu Abdurakhmonov, D. M. (2025). Fungal diseases of grapes and measures to combat them. Academic Journal of Science, Technology and Education, 1(5), 43-45.
66. qizi Sultonmurotova, S. M. (2025). Effectiveness and development prospects of digital-pedagogical integration in English language education. Academic Journal of Science, Technology and Education, 1(5), 16-20.
67. Rajabov, S. B., & ogli Alimov, J. S. (2025). Modern Approaches to Modernizing the Management System in Higher Education through Digital Technologies. Academic Journal of Science, Technology and Education, 1(5), 34-38.
68. Tolipov, M. S. (2025). The Relevance of Green Technology Integration in International Governance Processes. Academic Journal of Science, Technology and Education, 1(5), 3-6.
69. Xayrullo o'g'li, U. B., Khudoyberdiyev, B. S., & Xolmirzayev, M. M. (2025). The didactic potential of laboratory experiments in developing functional literacy. Academic Journal of Science, Technology and Education, 1(2), 50-54.