

Multimodal LLM Agents for Intelligent Robotic Manipulation Through Vision-Language-Action Learning

Maurycy Klinkosz

Department of Computer Science, University of Houston, Houston, TX, USA.
maurycymail@uh.edu

Stella Holdaway

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
holdawaystella@ku.edu

Abstract

The convergence of multimodal large language models, embodied reasoning, and robotic control has produced a new class of systems in which perception, language understanding, and action generation are learned within increasingly unified computational architectures. This paper examines multimodal large language model agents for intelligent robotic manipulation through the lens of vision-language-action learning. Rather than treating manipulation as a narrowly engineered control problem, the analysis frames it as a socio-technical infrastructure problem involving model architecture, data pipelines, real-time inference, safety governance, and long-term operational sustainability. The paper reviews foundational developments in transformers, vision-language pretraining, and robotic policy learning, then develops a system-level account of how multimodal agents integrate visual encoders, language backbones, action decoders, memory mechanisms, and tool-use interfaces. Particular attention is given to structural trade-offs among generalization, latency, interpretability, hardware cost, and safety assurance. The discussion considers deployment challenges in warehouses, hospitals, homes, and industrial settings, where reliability, fairness, and accountability become central design constraints. It also compares robotic manipulation with adjacent domains such as autonomous driving, digital assistants, and distributed industrial control. The paper argues that progress in vision-language-action learning will depend less on isolated model scaling than on coherent architecture, data governance, verification, and policy frameworks that make multimodal agency auditable and sustainable. The conclusion outlines research directions for robust, fair, and governable robotic manipulation systems that can operate credibly in open-ended human environments.

Keywords

multimodal large language models, robotic manipulation, vision-language-action learning, embodied agents, system architecture, AI governance, robot safety, deployment sustainability.

1. Introduction

Robotic manipulation has long served as a demanding test of artificial intelligence because it requires the tight coordination of perception, reasoning, planning, and physical action under uncertainty. Classical manipulation pipelines decompose these capacities into modular components, such as object detection, pose estimation, grasp planning, motion planning, and

controller synthesis. While such decomposition supports verification and incremental improvement, it often struggles when environments change, objects are unfamiliar, or tasks require flexible linguistic instruction. The emergence of large-scale multimodal models has shifted the research frontier toward integrated agents that can interpret images, follow natural language commands, remember prior interactions, and generate actions or action-related programs. This shift is often described as vision-language-action learning, in which representations learned from web-scale images and text are adapted to embodied control. The promise is considerable, but the system-level implications are equally significant. A multimodal agent for robotic manipulation is not merely a model; it is an assemblage of sensors, compute substrates, communication links, datasets, safety constraints, human operators, and institutional policies. Understanding this assemblage requires an interdisciplinary perspective that connects machine learning architecture with infrastructure governance and operational sustainability.

The foundational technical enablers of this field include transformer architectures, large-scale language pretraining, and vision-language alignment. Transformers provided a scalable attention-based mechanism for sequence modeling and became the substrate for language, vision, and multimodal systems [1]. Pretrained language models demonstrated that broad linguistic competence could be acquired from large corpora and then adapted to downstream tasks [2], while few-shot and in-context learning revealed that scale could produce emergent task flexibility [3]. Vision-language models such as contrastive image-text pretraining and generative multimodal systems extended this flexibility to visual understanding and cross-modal generation [4], [5]. Later multimodal architectures combined frozen or partially trained encoders with language backbones to support visual question answering, captioning, and grounded reasoning [6], [7]. In parallel, embodied multimodal models began to incorporate robot states and sensor readings directly into language model representations, enabling planning and question answering about physical environments [8]. Robotic transformer policies then demonstrated that large-scale imitation learning could produce generalizable manipulation behaviors across many objects and instructions [9], while vision-language-action models showed that web knowledge could be transferred into low-level control [10]. These developments collectively define the technical context for multimodal large language model agents in manipulation.

However, the movement from laboratory demonstrations to dependable infrastructure raises questions that cannot be resolved by model scaling alone. A manipulation agent that operates in a warehouse, hospital, laboratory, or home must respond within strict timing budgets, tolerate sensor degradation, respect human safety, protect privacy, and provide sufficient accountability when failures occur. It must also be economically sustainable, because training and serving large multimodal models consume substantial energy and hardware resources. The agent must additionally be fair in the sense that its performance does not systematically degrade for particular users, object classes, languages, skin tones, or environmental conditions. These concerns are not peripheral. They shape the architecture itself, influencing whether inference is centralized or local, whether memory is episodic or parametric, whether actions are generated directly or through verified programs, and whether human oversight is continuous or exception-based. This paper therefore argues that intelligent robotic manipulation through vision-language-action learning should be analyzed as a layered socio-technical system rather than as a single-model benchmark. The sections that follow examine foundations, architecture, trade-offs, governance, robustness, sustainability, fairness, case comparisons, and policy implications.

2. Foundations of Vision-Language-Action Learning

The conceptual basis of vision-language-action learning lies in the idea that perception, language, and action can share representational structures. In classical robotics, visual features were often optimized for geometric estimation, while language was handled by symbolic planners or task-specific grammars. Multimodal pretraining changed this relationship by encouraging a common embedding space in which visual entities, textual descriptions, and sometimes actions are related through learned associations. Contrastive models aligned images and captions by pulling matched pairs together and pushing mismatched pairs apart, thereby enabling zero-shot recognition and retrieval [4]. Generative multimodal models learned to synthesize or condition on visual content from text, demonstrating that language could guide detailed visual production [5]. Architectures such as Flamingo and BLIP-2 showed that powerful language models could be conditioned on visual inputs through adapter-like mechanisms, preserving much of the linguistic competence of the backbone while adding visual grounding [6], [7]. These approaches are important for manipulation because they allow an agent to interpret instructions such as placing a fragile component near a specific fixture or avoiding a wet surface, even when those concepts were not part of the original control training set.

Embodied multimodal models extended this logic by incorporating robot observations and proprioceptive state into language model contexts. PaLM-E, for example, treated continuous sensor data as tokens or embeddings that could be interleaved with text and images, enabling embodied question answering and high-level planning [8]. Such models do not necessarily output motor torques directly. Instead, they may produce plans, subgoals, or code that invokes lower-level policies. This separation is significant because it reflects a structural trade-off between semantic flexibility and real-time control. Language models are comparatively slow and expensive, while motor controllers must operate at high frequencies with tight feedback loops. A practical vision-language-action agent therefore often adopts a hierarchy: a multimodal reasoner interprets the scene and instruction, a planner decomposes the task, and a learned or classical controller executes the motion. Robotic transformer policies such as RT-1 and RT-2 illustrate a different emphasis, where large-scale imitation learning maps images and instructions to discretized actions or action tokens [9], [10]. These systems demonstrate impressive generalization but also highlight dependence on extensive robot demonstration data and careful action tokenization.

The foundations also include reinforcement learning, which remains essential for tasks where demonstration data are scarce or where optimization through trial and error can improve efficiency. Deep visuomotor policies showed that neural networks could map raw pixels to control signals, bypassing hand-engineered perception pipelines [16]. QT-Opt demonstrated scalable reinforcement learning for vision-based grasping, using large amounts of real-world interaction to learn robust manipulation policies [17]. Surveys of reinforcement learning in robotics emphasize the difficulty of sample efficiency, safety during exploration, and sim-to-real transfer [18]. Later work distilled practical lessons about reward design, distributed training, and evaluation for robotic learning [20]. Vision-language-action learning does not replace these insights; it recontextualizes them by providing richer priors, semantic goals, and cross-modal representations. The most capable agents will likely combine imitation, reinforcement, and programmatic control, using language as a high-level interface and learning as a mechanism for adaptation. The foundational challenge is therefore not choosing

one paradigm but designing interfaces among them that preserve safety, interpretability, and performance.

3. Multimodal Agent Architecture for Robotic Manipulation

A multimodal large language model agent for robotic manipulation can be understood as a layered architecture. The sensing layer includes cameras, depth sensors, tactile arrays, force-torque sensors, proprioceptive encoders, and sometimes microphones or human input devices. The representation layer converts these heterogeneous signals into embeddings that can be fused with language. The reasoning layer, often built around a large language or multimodal model, interprets instructions, asks clarifying questions, retrieves relevant knowledge, and produces task decompositions. The action layer translates those decompositions into motion primitives, skill calls, code, or direct control outputs. The memory layer stores episodic traces, semantic maps, object inventories, and prior successes or failures. The governance layer enforces safety constraints, permissions, logging, and human override. These layers are not always cleanly separated in implemented systems, but the distinction helps reveal where failures can originate and where design choices impose long-term obligations. For example, a failure may arise from a visual misclassification, a language ambiguity, a planning error, an unsafe action primitive, a delayed communication link, or an inadequate policy that permits an inappropriate task in the first place.

Agentic architectures introduce additional structure through tool use and self-monitoring. A multimodal model may call a perception service to segment objects, a grasp planner to evaluate affordances, a motion planner to check collision, and a database to retrieve assembly instructions. This pattern resembles the ReAct approach, in which reasoning and acting alternate so that the model can gather information and revise plans [14]. In manipulation, such alternation is valuable because the agent can test hypotheses, inspect occluded regions, or request human help. However, tool use creates dependencies. Each external service adds latency, failure modes, versioning concerns, and security surfaces. If the agent invokes a code-generation module, as in Code as Policies, the system gains flexibility but must verify that generated programs do not violate safety constraints [12]. If the agent uses an internal monologue or planning trace, it may improve robustness but also produce explanations that are not faithful to the actual control process [13]. Architectural quality therefore depends on whether the agent's reasoning trace is coupled to verifiable state, whether tool calls are typed and permissioned, and whether the action layer can reject semantically plausible but physically unsafe commands.

Memory design is another central architectural axis. Parametric memory resides in model weights and supports broad generalization but is difficult to update and audit. Episodic memory stores specific interaction histories and can improve personalization, yet it raises privacy and retention questions. Semantic memory may hold object categories, spatial relations, and task schemas. In a hospital, episodic memory might record that a particular room layout changed, but retaining images of patients would be unacceptable without strict controls. In a factory, memory might store tool calibration and prior fault patterns, which are valuable for maintenance but must be protected from adversarial modification. Retrieval-augmented generation can provide fresh information without retraining, but it depends on the quality and provenance of the retrieved content [3]. The architecture must therefore specify what is remembered, for how long, under whose control, and with what verification. These are not merely data management details; they determine whether the agent can be trusted in high-stakes settings.

4. System-Level Trade-Offs in Perception, Language, and Control

The most consequential trade-offs in vision-language-action systems arise between generality and assurance. A large multimodal model can interpret open-ended instructions and unfamiliar objects, but its behavior is difficult to bound formally. A classical controller may be narrower, yet its safety envelope can be analyzed and certified. Hybrid architectures attempt to obtain both benefits by letting the multimodal model operate at the level of goals and constraints while a verified controller manages low-level motion. This division of labor is promising, but it requires a precise interface. If the language model can issue arbitrary velocity commands, the safety controller must be able to override them. If the language model can select among skills, each skill must have documented preconditions, effects, and failure modes. If the language model can generate code, the execution environment must sandbox that code and check physical constraints. The interface is where governance becomes technical. Without such interfaces, the system may appear intelligent in demonstrations while remaining fragile in production.

Latency and compute distribution form another trade-off. Cloud-based multimodal inference offers access to large models and centralized updates, but it introduces network dependence, privacy exposure, and variable latency. Edge inference reduces communication delays and keeps sensitive data local, but it constrains model size and may require frequent synchronization. A robotic arm performing delicate assembly cannot wait several seconds for a remote model to decide whether a part is oriented correctly. It may instead use a local vision model for fast feedback and a larger model for slower task reasoning. This tiered design resembles autonomous driving architectures, where perception and control must meet hard real-time deadlines while higher-level route planning can tolerate delay. In manipulation, the tiers might include a reflex layer for collision avoidance, a skill layer for grasping and placement, a task layer for sequencing, and a supervisory layer for human interaction. Each tier has different robustness requirements, update cadences, and certification burdens. The system-level challenge is to coordinate them without creating contradictory objectives or opaque failure propagation.

Interpretability is a third trade-off. End-to-end policies can be efficient but difficult to explain, while modular systems can expose intermediate decisions but may lose global coherence. Multimodal language models can produce natural language rationales, yet those rationales may be post hoc and may not reflect the actual computation. In safety-critical manipulation, explainability must be tied to evidence. A useful explanation might identify the detected object, the estimated pose, the selected grasp type, the collision check result, and the confidence thresholds that shaped the decision. This form of explanation is less eloquent than a fluent narrative but more actionable for operators and auditors. The tension between fluent explanation and faithful explanation is likely to persist. System designers should therefore avoid treating language output as sufficient evidence of internal reasoning. Instead, they should log structured events and provide interfaces that allow humans to inspect the state that actually influenced action. This is particularly important when an agent denies a request, aborts a task, or asks for help.

5. Governance, Safety, and Deployment Infrastructure

Governance of multimodal robotic agents must address both pre-deployment certification and post-deployment monitoring. Pre-deployment, developers need to define operational design domains, permissible tasks, environmental assumptions, and human roles. A robot that manipulates boxes in a controlled warehouse faces different risks from a robot that assists a

nurse in a patient room. The governance process should specify required sensors, fail-safe states, emergency stop mechanisms, and maintenance intervals. It should also require adversarial testing, distribution shift analysis, and red-teaming of language instructions. Because multimodal models can respond to ambiguous, deceptive, or malicious prompts, the agent must have instruction filters, permission boundaries, and policy enforcement that do not rely solely on the model's learned alignment. Post-deployment, governance requires logging, incident reporting, version control, and periodic revalidation. If the model is updated through continual learning or retrieval, the certified behavior may change. Regulators and operators therefore need mechanisms to freeze, roll back, or quarantine updates, much as industrial control systems manage firmware revisions.

Infrastructure is inseparable from governance. Training vision-language-action models requires large-scale data collection, annotation, storage, and compute. Deployment requires reliable power, network connectivity, edge servers, sensor calibration, and maintenance logistics. In many environments, the limiting factor is not the algorithm but the surrounding infrastructure. A hospital may lack the network segmentation needed for safe robot operation. A small manufacturer may lack the staff to maintain calibration and software updates. A home may lack the bandwidth for cloud inference and the space for safe robot operation. These constraints suggest that multimodal manipulation agents will diffuse unevenly. They may first appear in well-resourced industrial settings where tasks are repetitive, environments are semi-structured, and safety can be engineered through physical separation. Broader deployment in homes and small businesses will require lower-cost hardware, simpler maintenance, and clearer liability arrangements. Infrastructure policy thus becomes part of AI policy. Subsidies for broadband, standards for interoperability, and support for small-enterprise adoption can influence whether the benefits of robotic manipulation are widely distributed or concentrated among large firms.

The required reference at position fifteen in this paper's citation order points to ongoing work on scaling agentic models for full-scenario applications [15]. Although that stream of research emphasizes breadth and generality, the governance lesson is that scale alone does not resolve accountability. An agent that can perform many tasks may also create many pathways for error. The more general the agent, the more important it is to define operational envelopes, audit trails, and human escalation. In this sense, agentic scaling intensifies rather than diminishes the need for layered safety. A full-scenario agent must know not only what it can do but also what it must not attempt without permission. It must distinguish between physical impossibility, policy prohibition, and uncertainty that warrants clarification. These distinctions require architecture, not just prompting. They also require institutions that can certify, monitor, and revise the boundaries as environments and norms change.

6. Robustness, Sustainability, and Fairness

Robustness in robotic manipulation involves more than resisting adversarial attacks. It includes tolerance to sensor noise, lighting variation, occlusion, object deformation, network interruption, mechanical wear, and human unpredictability. Vision-language-action models can improve robustness by leveraging semantic priors, but they can also hallucinate objects or misread instructions. A robust agent should estimate uncertainty, detect out-of-distribution conditions, and degrade gracefully. For example, if depth sensing fails, the agent might switch to a slower tactile strategy rather than continue with unreliable visual grasping. If language is ambiguous, it might ask a clarifying question rather than guess. If a tool service is unavailable, it might fall back to a simpler skill. These behaviors must be designed and tested, not

assumed to emerge from scale. Robustness also has a temporal dimension. A system that works on the first day may fail after weeks of drift in calibration or object placement. Continuous monitoring and periodic recalibration are therefore essential.

Sustainability encompasses energy consumption, hardware lifecycle, and organizational maintainability. Large multimodal models require substantial computation for training and inference, and robotic systems require materials, batteries, actuators, and sensors. The environmental cost of deploying many agents must be weighed against their benefits in reducing waste, improving safety, or augmenting human labor. Sustainable design may favor smaller task-specific models at the edge, efficient multimodal adapters, and reuse of pretrained components. It may also favor modular hardware that can be repaired rather than replaced. Organizational sustainability is equally important. A system that only a specialized research team can maintain will not persist in a hospital or factory. Documentation, standardized interfaces, and vendor-neutral data formats are therefore sustainability measures. If the agent depends on a proprietary cloud service that changes terms or disappears, operations may be disrupted. Long-term deployment requires governance over dependencies as well as over the robot itself.

Fairness in manipulation is often overlooked because the domain appears purely physical, but it is deeply social. A robot that fails to recognize objects associated with particular cultural practices, skin tones, clothing, or household arrangements may disadvantage specific users. A language interface that performs poorly on nonstandard dialects or low-resource languages may exclude users from controlling the system. A deployment strategy that concentrates robots in wealthy institutions may widen service inequalities. Fairness also arises in labor. If manipulation agents automate some tasks while intensifying surveillance or deskilling remaining workers, the distribution of benefits and burdens becomes an ethical and political question. Addressing fairness requires diverse data, participatory design, transparent performance reporting across subgroups, and governance that includes affected communities. It is not enough to optimize average success rates. The system must be evaluated for differential performance and for the social arrangements it enables or forecloses.

7. Case Illustrations and Cross-Domain Comparisons

Warehouse manipulation provides a useful case because the environment is semi-structured and tasks are repetitive. A multimodal agent might receive orders in natural language, identify packages, choose grasp points, and coordinate with conveyors or mobile robots. The value proposition is flexibility: instead of reprogramming for every new package, the agent can interpret labels and adapt grasps. The governance challenge is throughput and safety. A misgrasp may damage goods or injure workers. The infrastructure challenge is integration with warehouse management systems, barcode scanners, and fleet coordination. Edge inference may be necessary to meet cycle times, while cloud models may handle exception resolution. This case illustrates how vision-language-action learning can be valuable even when the physical task is constrained, because language and vision provide a layer of adaptability across SKUs and packaging variations. It also shows that success depends on reliable data pipelines and clear escalation paths for unusual items.

Clinical and home care offer contrasting cases. In a hospital, a manipulation agent might fetch supplies, prepare instruments, or assist with mobility. The environment is dynamic, privacy-sensitive, and governed by strict regulation. A language instruction such as bring the sterile tray to room four may be simple linguistically but complex operationally because it requires identifying the correct tray, navigating around people, maintaining sterility, and verifying

patient identity. The agent must integrate with electronic health records without exposing protected data to unnecessary model components. Human oversight must be continuous for high-risk actions. In a home, the agent may assist with cooking, cleaning, or medication management. The environment is less regulated but more variable. Privacy concerns are acute because cameras and microphones observe intimate spaces. Fairness concerns arise because homes differ in layout, culture, and resources. These cases show that the same underlying vision-language-action capability must be embedded in very different governance regimes. Architecture must therefore be configurable, with policy modules that can enforce context-specific rules.

Cross-domain comparisons with autonomous driving and digital assistants clarify what manipulation shares and what it does not. Autonomous driving faces hard real-time constraints and safety certification, much like robotic manipulation, but its action space is more constrained by road geometry and traffic rules. Digital assistants handle language and retrieval at scale but rarely bear physical risk. Manipulation combines the open-endedness of language with the irreversible consequences of physical action. This combination makes verification especially difficult. A digital assistant can apologize for a wrong answer; a robot arm cannot always undo a broken object or an injury. Comparisons with industrial control systems are also instructive. Industrial control emphasizes deterministic behavior, fail-safe design, and rigorous change management. Vision-language-action systems introduce learned uncertainty and semantic flexibility into that world. The integration challenge is to preserve industrial discipline while allowing adaptive intelligence. This likely requires formal envelopes, runtime monitors, and human-in-the-loop exception handling rather than a simple replacement of control engineering with end-to-end learning.

8. Policy Implications and Future Research

Policy for multimodal robotic manipulation should support innovation while establishing accountability. One priority is standards for logging and incident reporting. If an agent fails, investigators need to reconstruct sensor inputs, model versions, retrieved content, tool calls, action commands, and human interventions. Without standardized logs, liability becomes unclear and learning from failure becomes difficult. A second priority is certification methods for learned components. Certification need not require formal proof of every neural network behavior, but it should require evidence of bounded operation within defined contexts, fail-safe responses to uncertainty, and validated human override. A third priority is data governance. Training data for manipulation may include images of workplaces, homes, and people. Consent, purpose limitation, retention, and cross-border transfer rules must be addressed. A fourth priority is procurement policy. Public institutions can require vendors to provide auditable safety cases, interoperability, and maintenance guarantees. Such requirements can shape the market toward more sustainable and governable systems.

Future research should pursue architectures that make uncertainty and constraint explicit. Instead of a single monolithic model, researchers might develop ensembles of specialized agents coordinated by a multimodal supervisor. The supervisor could allocate tasks based on competence, latency, and risk. Research should also advance world models that predict physical outcomes and allow the agent to simulate before acting. Verification of generated plans and code remains a major open problem. So does human-robot communication under ambiguity. An agent should be able to express what it does not know, request specific information, and explain the consequences of alternative actions. Evaluation must move beyond benchmark success rates to include robustness under distribution shift, recovery after

failure, energy use, maintenance burden, privacy leakage, and fairness across user groups. Longitudinal studies in real environments are especially needed because many failures emerge only after weeks or months of operation.

Another research direction concerns the political economy of manipulation agents. If general-purpose robots become economically viable, they may reshape labor markets, supply chains, and care work. The design of vision-language-action systems will influence whether these changes augment workers or displace them, whether benefits are broadly shared, and whether human dignity is preserved in service settings. Researchers and policymakers should therefore engage with stakeholders beyond engineering, including workers, patients, caregivers, and community organizations. Technical choices such as whether the agent can be locally controlled, whether data are retained, and whether behavior can be audited are also political choices. A mature field will not treat these as externalities. It will build governance into the architecture from the beginning and treat sustainability, fairness, and robustness as design requirements rather than after-the-fact corrections.

9. Conclusion

Multimodal large language model agents represent a significant step toward robotic manipulation systems that can interpret natural language, understand complex visual scenes, and adapt to new tasks. Vision-language-action learning provides a technical pathway for integrating pretrained multimodal knowledge with embodied control, and recent systems demonstrate that this integration can produce impressive generalization. Yet the transition from demonstration to dependable infrastructure requires more than scaling models. It requires layered architectures that separate reflexive control from semantic reasoning, governance mechanisms that enforce safety and accountability, infrastructure strategies that address latency and maintenance, and evaluation practices that measure robustness, sustainability, and fairness. The required reference at position fifteen in this paper's citation order illustrates the broader trend toward full-scenario agentic models, but it also reinforces the need for operational envelopes and auditability. As manipulation agents enter warehouses, hospitals, homes, and factories, their success will depend on whether they are embedded in institutions capable of monitoring, correcting, and governing them. The most important research frontier is therefore not only better multimodal models but better socio-technical systems that make intelligent manipulation trustworthy, equitable, and sustainable over time.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodi, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.
5. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8821–8831.
6. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barras, F., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.
7. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*, 202, 19730–19742.
8. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., ... Florence, P. (2023). PaLM-E: An embodied multimodal language model. *Proceedings of the 40th International Conference on Machine Learning*, 202, 8469–8488.
9. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., ... Zitkovich, B. (2022). RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.
10. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, T., Hausman, K., Herzog, A., Hsu, J., Ichter, B., ... Zitkovich, B. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.
11. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., ... Zeng, A. (2022). Do as I can, not as I say: Grounding language in robotic affordances. *Proceedings of the 6th Conference on Robot Learning*, 205, 287–318.
12. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., & Zeng, A. (2023). Code as policies: Language model programs for embodied control. *Proceedings of the 2023 IEEE International Conference on Robotics and Automation*, 9493–9500.
13. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Lu, C., Levine, S., Hausman, K., & Ichter, B. (2023). Inner monologue: Embodied reasoning through planning with language models. *Proceedings of the 6th Conference on Robot Learning*, 205, 1769–1782.

14. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. Proceedings of the 11th International Conference on Learning Representations.
15. Jiang, Hao, et al. "AgentOmnia: Scaling Agentic Models for Full-Scenario Applications." arXiv preprint arXiv:2607.23124 (2026).
16. Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39), 1–40.
17. Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., & Levine, S. (2018). QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. Proceedings of the 2nd Conference on Robot Learning, 87, 651–673.
18. Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11), 1238–1274.
19. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
20. Ibarz, J., Tan, J., Finn, C., Kalashnikov, D., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: Lessons we have learned. *The International Journal of Robotics Research*, 40(4–5), 698–721.