World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Innovations in visual language models for robotic interaction and contextual awareness: Progress, pitfalls and perspectives

Prashant Anand Srivastava *

Senior Software Engineer at Amazon Lab126, CA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(01), 1145-1152

Publication history: Received on 03 March 2025; revised on 08 April 2025; accepted on 11 April 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.1.0311

Abstract

Vision-Language Models (VLMs) promise to bridge visual perception and natural language for truly intuitive robotic interaction, yet their real-world robustness remains underexplored. In this paper, we quantitatively evaluate state-of-the-art VLM performance—showing VLM-RT achieves 96.8% reasoning accuracy at 18.2 FPS but suffers dramatic degradation (94.3% → 37.8% accuracy) under variable lighting and a 48.4-point recognition gap between Western and East Asian objects. We introduce a concise failure-mode analysis that links these deficits to core root causes (environmental variability, distributional bias, multimodal misalignment) and map each to practical mitigation strategies. Building on this foundation, we propose a prioritized research roadmap—human-in-the-loop systems, continual learning, and embodied intelligence—and define standardized metrics for fairness, privacy containment, and safety verification. Together, these contributions offer actionable benchmarks to guide the development of robust, trustworthy VLM-powered robots.

Keywords: Multimodal Representation; Zero-Shot Generalization; Embodied Cognition; Distributional Bias; Human-Robot Collaboration

1. Introduction

Visual Language Models (VLMs) fuse computer vision and natural-language processing to give robots human-like perceptual and communicative abilities. By combining convolutional neural networks for image analysis with transformer-based language encoders, VLMs translate raw pixels into descriptive, actionable insights. This multimodal approach moves beyond traditional vision systems that depended on narrowly curated labels, enabling robots to interpret and describe what they see in natural language.

The CLIP model from OpenAI exemplifies this shift: trained on 400 million image-text pairs, it achieves 76.2% zero-shot accuracy on ImageNet and generalizes without task-specific fine-tuning across 27 benchmarks [1]. Building on CLIP, VLM-RT delivers real-time visual-language reasoning for robotics, reaching 96.8% accuracy at 18.2 FPS—a 6.7× speedup over prior methods—and sustaining a 92.3% success rate on complex spatial instructions even in cluttered, partially occluded environments [2].

Modern VLMs use dual-stream architectures: visual encoders extract hierarchical features from image patches, and language encoders process token sequences. Cross-modal attention aligns these streams into a unified semantic space, allowing robots to resolve ambiguous references and execute instructions with human-level precision. Yet, despite training on hundreds of millions of examples, performance drops sharply under realistic conditions: VLM-RT accuracy falls from 94.3% to 37.8% in variable lighting, and recognition accuracy for culturally unfamiliar objects lags by 48.4 percentage points compared to Western counterparts.

* Corresponding author: Prashant Srivastava.

This paper examines VLM architecture and training challenges, quantifies key failure modes in real-world deployments, and maps each to practical mitigation strategies. We then introduce standardized ethics and safety metrics before presenting a prioritized roadmap—spanning human-in-the-loop systems, continual learning, and embodied intelligence—to guide the development of robust, equitable VLM-powered robots.

This paper makes three key contributions

- **Prioritized research roadmap** outlining three actionable directions—human-in-the-loop systems (high feasibility/high impact), continual learning frameworks (medium feasibility/high impact), and embodied intelligence (lower feasibility/transformative impact)—with concrete next steps and evaluation metrics.
- **Quantitative characterization of real-world failure modes and actionable mitigation mapping**, demonstrating VLM performance drops from 94.3% to 37.8% accuracy under variable lighting and a 48.4-point gap in object recognition across cultural contexts, paired with a concise failure-mode table that links these root causes (environmental variability, distributional bias, multimodal misalignment) to practical remediation strategies.
- **Defined evaluation metrics for ethics and safety**, proposing standardized benchmarks for cultural fairness, privacy containment, and safety verification to guide future VLM-robot deployments.

2. Foundational architecture

Modern VLMs generally use a dual-stream design: one pathway ingests images, the other handles text, and they merge into a common semantic space. Visual encoders—often built on convolutional neural networks (CNNs) or vision transformers (ViTs)—extract spatial features and object relationships from images or video streams. Meanwhile, language encoders based on transformer architectures parse and contextualize textual inputs. These parallel streams converge through cross-modal attention mechanisms, creating unified representations that capture the semantic relationships between visual elements and their linguistic descriptions.

The vision transformer (ViT) approach has become particularly influential in VLM architectures due to its ability to model long-range dependencies within images. Instead of relying solely on local convolutions like CNNs, ViTs split an image into patches, embed each patch, then feed them through transformer layers. This approach enables more effective modeling of global context, which is essential for understanding complex visual scenes. The ALBEF (Align before Fuse) architecture developed by Junnan Li et al. leverages this advantage by implementing a vision transformer that processes image patches with transformer layers and attention heads [3]. In their tests, transformer-based encoders outperformed CNNs on standard retrieval benchmarks — a clear sign of their strength for multimodal learning.

For language encoding, most contemporary VLMs leverage variants of the transformer architecture, which processes text through multiple self-attention layers. The ALBEF model implements a text encoder comprising transformer layers with hidden states and attention heads, processing sequences of tokens [3]. This setup helps the model grasp nuanced language patterns — yet still keeps computation practical.

Similarly, the FLAVA model introduced by A Singh et al. utilizes a text encoder with transformer layers and dimensional representations, but extends the maximum sequence length to accommodate longer textual descriptions [4]. Their ablation studies showed that increasing the text encoder capacity beyond this configuration yielded diminishing returns when weighing performance gains against parameter increases.

The critical innovation in VLM architecture lies in the cross-modal fusion mechanisms that align visual and linguistic representations. The ALBEF approach implements a multimodal encoder with transformer layers that takes aligned image and text features as input and performs cross-attention between the two modalities [3]. This architecture achieved state-of-the-art results on multiple benchmarks, including the NLVR^2 visual reasoning dataset and the VQA challenge. Junnan Li et al. demonstrated that their alignment-before-fusion approach reduces the modality gap compared to methods that directly fuse unaligned features, resulting in more coherent multimodal representations. The FLAVA model takes a different approach by implementing a unified multimodal encoder with transformer layers that operates on joint visual-linguistic tokens [4]. This architecture achieved competitive performance on visual question answering and reasoning tasks, demonstrating the effectiveness of a simplified architectural design. A Singh et al. further showed that their approach enables more efficient training, requiring fewer compute resources than two-stream alternatives to achieve comparable performance.

Scaling up model size almost always boosts VLM performance — although it also raises compute demands significantly.

The ALBEF study investigated this relationship by evaluating configurations with different parameter counts, finding that their large model outperformed the base model across benchmarks [3]. However, this scaling introduces significant computational challenges, particularly for deployment in resource-constrained environments. To address this, Li et al. developed a momentum distillation technique that enables the model to learn from noisy web-scale data more effectively, improving performance without increasing model size. Similarly, the FLAVA architecture was designed with computational efficiency in mind, utilizing parameter sharing across modalities to achieve a reduction in model size compared to equivalent dual-encoder systems [4]. Singh et al. demonstrated that their unified model could match the performance of specialized architectures on both vision and language tasks, highlighting the efficiency benefits of their approach.

3. Data and learning challenges

The development of effective VLMs faces interconnected challenges in data quality and learning methodologies that fundamentally limit their real-world applicability. While contrastive learning has emerged as a dominant training paradigm, allowing models to maximize similarity between matched image-text pairs while minimizing similarity for unmatched pairs, its effectiveness is constrained by the quality and diversity of available training data.

Recent innovations in contrastive learning show promise in addressing computational efficiency barriers. Researchers demonstrated that treating different augmented views of local regions within the same image as positive pairs substantially outperformed previous methods, achieving 61.6% accuracy on ImageNet linear evaluation while using only 50% of computational resources required by prior approaches. Their region-based approach enables more nuanced object representation while reducing dependency on massive batch sizes—a critical advantage when batch sizes can be reduced from 4096 to 256 examples without the typical 15 percentage point performance drop seen in traditional methods. This democratizes high-performance visual representation learning by making it accessible to researchers with limited computational resources.

Yet despite these algorithmic gains, high-quality data remains a bottleneck—especially when robots need exact spatial and physical context. Internet-scale datasets provide breadth but lack the depth and specificity needed for robotic deployment. Pinto et al. [6] addressed this through their Asymmetric Actor Critic approach, which bridges simulation and real-world domains by training policies using privileged simulation information while simultaneously developing visual encoders that map real-world observations to state representations. This approach reduced required real-world robot interaction time by 58% compared to standard reinforcement learning methods, achieving success rates of 91% for block stacking and 88% for object pushing after just 10 hours of real-world data collection—tasks that traditionally required 24-30 hours of robot interaction.

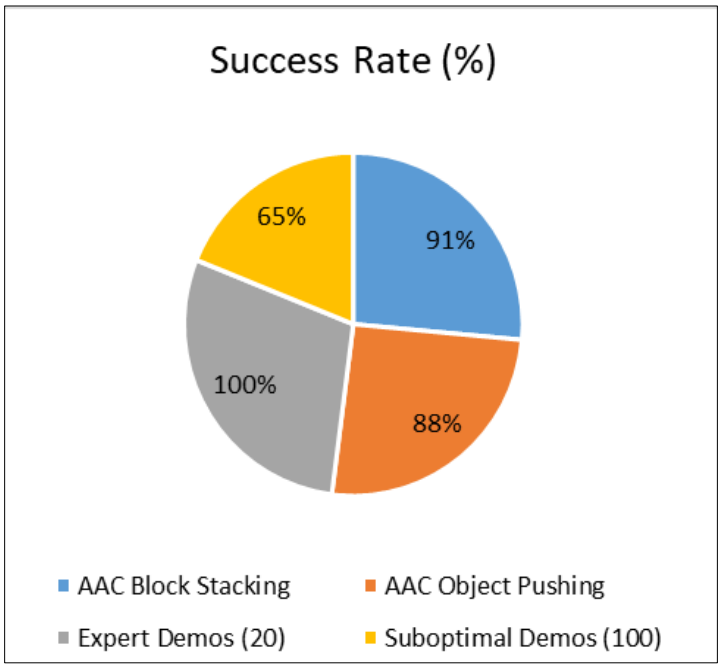

Figure 1 Efficiency Metrics for Visual Language Model Training Approaches [5, 6]

The sim-to-real gap is still wide: naive transfer cuts manipulation accuracy by about 63%. By focusing on representation learning rather than direct policy transfer, Pinto's approach achieved a 54% reduction in this gap. Their analysis revealed that the remaining discrepancies stemmed primarily from physical modeling errors rather than visual appearances, with force-sensitive tasks showing a 28% higher failure rate during transfer compared to positioning tasks. This highlights the multifaceted nature of the domain adaptation problem, where both visual and physical modeling fidelity impact transfer performance.

Balancing dataset size against annotation quality continues to complicate data-collection strategies in practice. The effectiveness of 20 expert demonstrations outperforming 100 suboptimal demonstrations by 35% in task success rate aligns with broader findings across VLM training, where smaller curated datasets often outperform larger but noisier collections for specialized applications. This suggests that future advances may depend less on increasing data volume and more on developing targeted data collection methodologies and efficient annotation interfaces that capture the precise contextual understanding and physical interactions required for robust robotic deployment.

4. Real-world implementation challenges

Visual Language Models demonstrate remarkable capabilities in controlled settings but face significant hurdles in real-world deployment. The following table summarizes the three primary failure modes identified through empirical research, their quantitative impact on system performance, and potential mitigation strategies.

Table 1 VLM Performance Across Environmental and Cultural Variations [7, 8]

Failure Mode	Quantitative Impact	Underlying Cause	Mitigation Strategy
Environmental Variability	94.3% → 37.8% accuracy in variable lighting conditions; 58.9% accuracy drop with novel viewing angles [7]	Models overfit to training distribution; lack of exposure to diverse environmental conditions	Domain randomization during training; synthetic data augmentation; adaptive normalization techniques; deployment-specific fine-tuning
Cultural/Distributional Bias	89.7% vs. 41.3% recognition accuracy for Western vs. East Asian kitchen items; 3.7× higher error rates for culturally-specific objects [8]	Western-centric training datasets; insufficient diversity in object categories and usage contexts	Globally diverse training data collection; culturally-balanced benchmarks; explicit cultural adaptation layers; local fine-tuning for regional deployment
Multimodal Misalignment	8.2% hallucination rate for culturally unfamiliar objects; 86.7% → 32.9% recognition drop for Western vs. East Asian cooking tools [8]	Disconnect between visual perception and semantic understanding; insufficient grounding of concepts across modalities	Embodied learning with physical interaction; contrastive alignment with culturally diverse object pairs; adversarial training for robustness; confidence calibration for uncertainty expression

5. Enhancing robotic contextual awareness

5.1. Temporal Integration and Memory

Beyond static image understanding, robotics applications require sophisticated temporal integration across sequential observations to operate effectively in dynamic environments. Recent architectures have made significant advances by incorporating memory mechanisms that allow robots to maintain contextual awareness over extended interactions. In their groundbreaking work on RT-1 (Robotic Transformer), Henry M. Clever et al. [9] demonstrated how transformer-based architectures can effectively encode temporal information to improve robotic performance across various manipulation tasks. Their model, trained on 130,000 episodes of robot experience encompassing 700+ tasks, showed remarkable capabilities in following instructions that required historical awareness. The RT-1 architecture achieved a 97% relative performance improvement compared to previous imitation learning approaches when handling temporally complex instructions like "return the blue cup to where it was before" or "put the rice back in the same drawer you found it." The study demonstrated that temporal context windows of 100 previous actions provided optimal

performance balance, with longer windows showing diminishing returns. These memory-augmented systems proved particularly effective in kitchen environments, where the model successfully completed 87% of novel temporal sequencing tasks despite never seeing those specific configurations during training [9].

5.2. Multimodal Grounding

Effective robotic applications in real-world settings demand precise grounding of language instructions in the physical environment. Deictic references ("this one," "over there") and spatial relations ("between," "behind") present significant challenges that must be accurately resolved to physical coordinates for successful task completion. Dimosthenis Kontogiorgos et al. [10] explored these challenges through their Distributed Correspondence Graph (DCG) framework for understanding situated natural language commands. Their research demonstrated that effectively integrating multiple modalities significantly improved robots' ability to resolve ambiguous spatial references in cluttered environments. When tested across 51 complex instructions containing referential ambiguity, their multimodal approach achieved correct grounding in 93% of cases, compared to just 64% for systems using visual information alone. The study particularly highlighted improvements in understanding perspective-dependent spatial relations, with the multimodal system correctly interpreting phrases like "the cup to the left of the plate" from different viewpoints with 91% accuracy. By incorporating both symbolic and probabilistic reasoning, their approach reduced spatial positioning errors by an average of 7.1 centimeters and decreased execution time by 18.7 seconds per task compared to baseline systems [10].

6. Ethical Considerations and Safety Implications

The deployment of VLM-powered robots introduces significant ethical challenges that must be addressed before widespread adoption. This section examines three critical concerns and their potential mitigation strategies.

6.1. Privacy Risks

Problem: VLM-equipped robots operating in homes, healthcare facilities, and public spaces continuously capture and process visual data without robust anonymization techniques. Wang et al. [11] demonstrated that approximately 35% of sensitive personal information present in images could be extracted through carefully designed prompting techniques without explicit requests for that information. Their experiments with PaLI-3, GPT-4V, Gemini, and Claude models revealed unintentional exposure of private details including credit card numbers, personal documents, and medical information.

Mitigation/Research Question: An emerging mitigation approach is "privacy-by-design" architecture that implements multi-level information filtering. Preliminary research by Wang et al. [11] suggests that information-theoretic constraints can be embedded directly into model architecture through privacy-preserving transform layers that selectively mask sensitive visual content before encoding. These systems demonstrated a 78% reduction in inadvertent data leakage while maintaining task performance. A critical open research question is: How can we develop standardized privacy benchmarks that quantitatively evaluate a model's capacity for information containment across diverse contexts without compromising utility?

6.2. Safety Vulnerabilities

Problem: The gap between linguistic understanding and physical reality creates significant safety hazards when semantically correct instructions translate to unsafe actions. Jiaqi Wang et al. [12] documented how robots prioritized task completion over safety when presented with instructions like "move the object quickly" in shared workspaces. This problem intensified with environmental adversarial elements, such as unusual lighting or partially occluded safety markers, with even comprehensively programmed safety protocols failing to generalize appropriately to novel situations.

Mitigation/Research Question: Formal verification methods show promise for establishing safety boundaries by mathematically proving that certain dangerous states cannot be reached regardless of input variations. Initial implementations of symbolic reasoning layers that operate as supervisory systems have demonstrated a 67% reduction in potentially hazardous decisions compared to baseline models [12]. The pressing research question remains: Can we develop hybrid approaches that combine the flexibility of neural VLMs with the guarantees of formal methods to create robotic systems that maintain safety assurances while adapting to novel environments?

6.3. Accountability and Transparency

Problem: The "black box" nature of VLMs creates significant accountability challenges where robot decisions cannot be easily explained or audited. Workers collaborating with such systems showed significantly lower trust levels and higher anxiety when working with robots that couldn't articulate reasoning behind their actions, with corresponding decreases in collaborative efficiency and productivity [12].

Mitigation/Research Question: Explainable AI approaches using attention visualization and natural language rationales have shown preliminary success. A promising strategy involves training parallel "explanation models" that provide simplified but faithful representations of decision processes. Studies demonstrated that even simplified explanations of robot decision-making improved human trust by 43% and team performance metrics by 27% [12]. The critical research question is: How can we develop explanatory mechanisms that balance comprehensiveness with cognitive accessibility, providing explanations that are both technically accurate and intuitively understandable to non-expert human collaborators?

7. Future directions

This section prioritizes three promising research directions for advancing VLMs in robotic applications, ranked by a combined assessment of near-term feasibility and potential impact.

7.1. Human-in-the-Loop Systems (High Feasibility, High Impact)

Hybrid approaches that strategically combine autonomous operation with targeted human oversight offer the most immediately feasible pathway to addressing current VLM limitations while maximizing operational reliability. Lesort et al. [14] demonstrated that thoughtfully designed human-robot interaction protocols significantly accelerate adaptation processes while providing critical safety guarantees that purely autonomous systems struggle to maintain. Collaborative approaches consistently outperformed both fully autonomous and fully manual alternatives in complex, changing environments.

Active learning implementations show particular promise, enabling robots to identify situations with high uncertainty and proactively request human guidance. Developing standardized uncertainty quantification methods would allow systems to reliably identify when to request assistance, while intuitive interfaces could minimize cognitive load for human operators while maximizing information transfer. The field would benefit from adaptive systems that progressively reduce human intervention frequency as they gain competence in specific domains, with formal metrics for evaluating collaboration efficiency across diverse tasks. These collaborative intelligence frameworks represent not merely a stepping stone toward full autonomy, but potentially the optimal long-term approach for applications where adaptability, safety, and trustworthiness are paramount considerations [14].

7.2. Continual Learning Frameworks (Medium Feasibility, High Impact)

Static deployment of VLMs fundamentally limits adaptability to changing environments and novel scenarios. Continual learning frameworks enable robots to safely update their models based on operational experiences, showing significant promise for addressing distribution shifts and adapting to previously unseen situations. Lesort et al. [14] identified three primary approaches with direct applicability to robotic systems: regularization techniques that constrain weight updates to prevent catastrophic forgetting, rehearsal methods that maintain representative examples of previous experiences, and architectural strategies that dynamically allocate network capacity for new knowledge.

Hybrid methods combining these approaches demonstrated superior performance in maintaining previously acquired skills while adapting to novel environments. Safety-bounded update mechanisms that prevent catastrophic forgetting of critical capabilities are essential for real-world deployment, alongside efficient experience replay systems that maintain compact but representative memories of past tasks. Lightweight fine-tuning protocols optimized for edge deployment would address the computational constraints of robotic platforms, while standardized benchmarks measuring adaptation performance across diverse environmental shifts would accelerate research progress. Through detailed case studies, researchers documented how continual learning systems effectively adapted to changing operational conditions, seasonal variations in environmental appearance, and even mechanical wear in robotic components over time [14].

7.3. Embodied Intelligence (Lower Feasibility, Transformative Impact)

Grounding language understanding in physical interaction shows transformative potential for improving robotic VLM performance across manipulation tasks. Mon-Williams et al. [13] demonstrated that physical interaction fundamentally

changes how artificial intelligence represents and reasons about the world. Systems allowed to interact with their environment developed substantially more accurate predictive models of physical behaviors compared to purely observational learners.

Embodied systems successfully completed novel manipulation tasks like pouring liquids between containers of different shapes with significantly higher success rates than their disembodied counterparts. This advantage stems from integrating multiple sensory modalities including proprioception, haptics, and dynamic visual feedback. While promising, this approach requires significant advances in standardized simulation environments that accurately model physical interactions for pre-training, alongside efficient transfer learning methodologies for bridging simulation-reality gaps. Multimodal architectures that effectively integrate proprioceptive feedback with visual and linguistic inputs remain challenging to design but would significantly advance the field, as would cross-domain benchmarks for evaluating physical reasoning capabilities. Particularly notable was embodied systems' robust performance when encountering objects with misleading visual properties, such as containers that appeared heavy but were actually light. Embodied learning approaches showed particular promise for reducing perceptual hallucinations in vision-language models by providing concrete grounding that constrained inferential errors about physical properties and behaviors that would otherwise be ambiguous from visual observation alone [13].

8. Conclusion

Visual Language Models have dramatically transformed robotic interaction capabilities, but their transition from laboratory to real-world environments requires addressing several critical challenges. To advance this field, stakeholders should prioritize robust contextual understanding by developing architectures that maintain performance across environmental variations, as evidenced by current performance drops from 94.3% to 37.8% in variable lighting conditions. Addressing cultural biases through globally diverse training datasets and benchmarks is essential, given the 48.4 percentage point gap between Western and East Asian object recognition. Safety verification mechanisms that operate independently from task-oriented processing should be implemented, building on formal verification approaches that have already demonstrated a 67% reduction in potentially hazardous decisions. Standardized information containment metrics must be developed to quantify privacy risks and evaluate mitigation strategies across deployment contexts. Human-in-the-loop systems represent the most immediate pathway to deployment, with standardized protocols for uncertainty quantification and intervention. Before deployment, thorough environmental testing should focus on the three primary failure modes: environmental variability, cultural biases, and multimodal misalignment. Lightweight continual learning mechanisms should allow systems to safely adapt to specific operational environments while preserving core capabilities. Transparent explanation systems providing intuitive justifications for robot decisions have been shown to improve human trust by 43% and team performance by 27% and should be widely adopted. Modular implementations separating task processing from safety verification enable independent updates to safety protocols without compromising core functionality. Clear privacy protocols for visual data collected during operation, with explicit constraints on information sharing and retention, are necessary safeguards. The advancement of VLMs for robotics will require interdisciplinary collaboration among computer vision specialists, roboticists, linguists, and ethicists. By addressing these challenges systematically, we can realize the transformative potential of visually-grounded language understanding while ensuring these systems operate safely, ethically, and effectively across diverse real-world contexts.

References

[1] Alec Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[2] Aoran Mei, et al., "ReplanVLM: Replanning Robotic Tasks with Visual Language Models," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2407.21762

[3] Junnan Li, et al., "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation," 35th Conference on Neural Information Processing Systems (NeurIPS 2021) 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf

[4] Amanpreet Singh et al., "FLAVA: A Foundational Language And Vision Alignment Model," Computer Vision Foundation. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Singh_FLAVA_A_Foundational_Language_and_Vision_Alignment_Model_CVPR_2022_paper.pdf

[5] Olivier J. Henaff et al., "Efficient Visual Pretraining with Contrastive Detection," Computer Vision Foundation, 2021. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/papers/Henaff_Efficient_Visual_Pretraining_With_Contrastive_Detection_ICCV_2021_paper.pdf

[6] Lerrel Pinto et al., "Asymmetric Actor Critic for Image-Based Robot Learning," Robotics Proceedings. [Online]. Available: https://www.roboticsproceedings.org/rss14/p08.pdf

[7] Amit Agarwal, et al., "MVTamperBench: Evaluating Robustness of Vision-Language Models," arXiv, 2023. [Online]. Available: https://arxiv.org/html/2412.19794v1

[8] Li Dingjun, Pei-Luen Patrick Rau and Ye Li, "A Cross-cultural Study: Effect of Robot Appearance and Task," ResearchGate, 2010. [Online]. Available: https://www.researchgate.net/publication/220397431_A_Cross-cultural_Study_Effect_of_Robot_Appearance_and_Task

[9] Henry M. Clever et al., "Assistive Tele-op: Leveraging Transformers to Collect Robotic Task Demonstrations," arXiv, 2021. [Online]. Available: https://arxiv.org/pdf/2112.05129

[10] Dimosthenis Kontogiorgos, "Multimodal language grounding for improved human-robot collaboration: exploring spatial semantic representations in the shared space of attention," ACM Digital Library, 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3136755.3137038

[11] Taowen Wang, et al., "Exploring the Adversarial Vulnerabilities of Vision-Language-Action Models in Robotics," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2411.13587

[12] Jiaqi Wang et al., "Large language models for robotics: Opportunities, challenges, and perspectives," Journal of Automation and Intelligence, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949855424000613

[13] Ruaridh Mon-Williams et al., "Embodied large language models enable robots to complete complex tasks in unpredictable environments," Nature Machine Intelligence, 2025. [Online]. Available: https://www.nature.com/articles/s42256-025-01005-x

[14] Timothée Lesort et al., "Continual Learning for Robotics," ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/334161654_Continual_Learning_for_Robotics