Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37945

Generalization Robustness of Vision-Language-Action Models in Open-World Scenarios

Xu Zhou

School of Future Technology, South China University of Technology, Guangzhou, China
202530362658@mail.scut.edu.cn

Abstract. Vision-Language-Action (VLA) models consistently perform exceptionally well in highly controlled simulations, but they currently experience a considerable drop in their generalization capabilities for complex, unpredictable open-world scenarios. Furthermore, existing academic surveys entirely lack comprehensive, systematic failure-mode classification alongside deeply mechanistic root-cause analysis. To directly address this gap, this paper proposes a four-dimensional robustness framework of objects, environments, languages, and time. Three main architectures (Robotic Transformer 2 (RT-2), Open-source Vision-Language-Action model (OpenVLA), VoxPoser) are compared, and optimisation schemes are divided into data, architecture, training, and deployment levels with unified metrics. Multi-dimensional combined optimisation has achieved the best results (86.4% unseen object success rate, 77.3% ambiguous instruction completion rate, 68.9% long-sequence accuracy), but despite this, persistent cumulative temporal errors remain the main problem even after full optimisation. Ultimately, the presented framework systematically puts forward the critical research content and provides a highly organised path for the following model design and future deployment.

Keywords: Vision-Language-Action Models, Open World, Generalization Robustness, Embodied Intelligence, Robotic Manipulation

1. Introduction

Vision-Language-Action (VLA) models are leading the way in integrating large language models, multi-modal perception and robotic control. They have formed the foundation of the technical system for embodied intelligent robots. The three models can learn scene understanding and interpret natural language instructions to continuously predict sequences of actions. Robots can perform all kinds of handling operations according to the natural language instructions issued by people. Recent progress in pre-trained vision-language backbones and large-scale robot demonstration datasets has significantly improved task success rates in standardised laboratory and simulation environments. Although these models have achieved good results in the lab, they often fail when encountering novel objects, altered lighting, ambiguous language, and long-horizon tasks in open-world deployment due to narrow data distribution. Several surveys have provided an overview of VLA architectures, their history, and real-world deployment problems [1-3]. However, the existing surveys have some deficiencies: they focus mainly on architectural taxonomies and do

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37945

not systematically classify generalization failure modes or explore the underlying mechanistic root causes. To correct the above defects, this paper builds a four-dimensional failure classification system, identifies five main root causes, compares three representative architectures, and quantitatively evaluates four types of optimisation schemes with a unified metric.

2. VLA model architecture and representative frameworks

2.1. Three core modules of VLA models

The three modules in a typical VLA model are used to process visual observations and language instructions to generate robot actions. The visual perception module is given a single-frame RGB image from the robot's camera as input, uses pre-trained encoders such as SigLIP and DINOv2 to extract compact visual feature vectors, but lacks depth and dynamic environmental information because it is a single-frame input. The multi-modal language-image fusion module then unifies visual features and text embedding vectors in a shared feature space to build cross-modal correspondence between human language commands and scene visual information. Finally, the action sequence prediction module takes the fused multi-modal features as input and outputs continuous joint position sequences or discrete action tokens to physically drive the robot manipulator in a specific manner.

2.2. Representative VLA models and comparative analysis

Three representative VLA frameworks lead the open-world manipulation field and have different architectural philosophies. RT-2, developed by Google DeepMind, pioneers the end-to-end VLA paradigm by co-finetuning a frozen vision-language model on web-scale image-text data and robot demonstration data, achieving strong performance on simple open-ended language commands but suffering sharp degradation on unseen objects and complex multi-step tasks due to the absence of explicit geometric perception [4]. OpenVLA is an open-source unified pre-training framework that combines multi-platform manipulation datasets to improve cross-robot transferability, and although it serves as a standard baseline for open-world VLA research, it is still prone to cumulative action errors in long-horizon sequential tasks [5]. VoxPoser is a modular system that separates 3D geometric reconstruction from language-guided motion generation, and uses voxel-based spatial value maps to achieve excellent generalization to unseen object combinations and viewpoint changes [6]. However, as a complete system, it is dependent on an external LLM for language parsing and lacks end-to-end cross-modal alignment training, thus limiting its ability to interpret gradable or context-dependent language commands [7]. As shown in the horizontal comparison, end-to-end architectures perform well in language grounding but lack spatial reasoning; modular frameworks are environmentally robust at the cost of system complexity; and open-source models focus on transferability but do not address temporal error correction.

3. Classification of generalization robustness failures in open worlds

3.1. Object generalization failure

Kim and others used the OpenVLA model to perform experiments on the LIBERO benchmark with 120 unseen household objects [5]. The baseline's success rate dropped to 41.7% on unseen categories from 62.3%, and object novelty was identified as the main reason for failure. Fei and

DOI: 10.54254/2755-2721/2026.37945

others also found that new combinations of familiar objects resulted in a 15.3% drop due to a failure in reasoning about unpaired spatial relationships [8].

3.2. Environmental generalization failure

Huang and others tested VoxPoser in six light conditions ranging from 50 lux to 800 lux, and a drop in success rate of 21.4 percentage points occurred; thus, single-frame RGB perception is highly sensitive to changes in light [8]. Fang et al. also found that rotating the 30-degree camera viewpoint decreased the success rate of OpenVLA by 18.7%, and viewpoint invariance remains a problem [7].

3.3. Linguistic instruction generalization failure

Brohan and others reported that RT-2 had a 71.5% completion rate for single-step instructions but only a 48.3% completion rate for multi-step commands, and the understanding of compositional language dropped by 23.2% [4]. Fang et al. added vague modifiers such as "slightly" and "carefully", and found that OpenVLA's accuracy dropped by 19.4% compared with explicit commands because the model lacked fine-grained semantic grounding for gradable adjectives [7]. In short, the semantic-action alignment mechanism for unstructured natural language is still lacking.

3.4. Temporal sequence generalization failure

Temporal sequence generalization failure arises from the accumulation of small per-step errors over long-horizon chains. Each independently predicted action introduces a deviation that accumulates in subsequent steps, thus diverging from the target. Fan and others have investigated the skill chaining problem in long-horizon VLA tasks and found that errors in the initial subtasks accumulate and spread, resulting in severe failures as the sequence length increases [9]. Liu and others found that reinforcement learning fine-tuning reduced the per-step error by 12.3%, but it failed to eliminate cumulative drift entirely; thus, they believe the cause is the lack of built-in temporal feedback and error correction mechanisms [10].

4. Core root causes of weak generalization robustness

Before listing the specific reasons for each, this section first presents an organised map of the four-dimensional generalization failure modes identified in Section 3 to the five main root causes analysed below. Table 2 shows this phenomenon-root-cause mapping matrix.

Table 1. Phenomenon-root cause mapping matrix for VLA open-world generalization failures

Root Cause \ Failure Mode	Object Generalization Failure	Environmental Generalization Failure	Linguistic Instruction Generalization Failure	Temporal Sequence Generalization Failure
Single-frame Visual Perception Limitation	● Strong	● Strong	○ Weak	○ Weak
Semantic-action Domain Gap	—	—	● Strong	◐ Moderate

DOI: 10.54254/2755-2721/2026.37945

Table 1. (continued)

Lack of Temporal Error Correction	—	—	—	● Strong
Absence of Physical Prior Constraints	◐ Moderate	○ Weak	—	◐ Moderate
Real-world Data Scarcity and Domain Shift	● Strong	● Strong	● Strong	● Strong

Note: ● Strong = direct primary causal factor; ◐ Moderate = contributing factor that increases the severity of failure; ○ Weak = indirect secondary influence; — = no direct causal relationship.

The matrix shows that real-world data scarcity is a general severe constraint for all four failure types, single-frame visual perception limitations mainly affect object and environment generalization, and the lack of temporal error correction exclusively causes temporal sequence failure. Three structural Patterns appear in the mapping matrix. First, no single root cause corresponds to a particular failure mode; therefore, open-world generalization degradation is a multi-factor problem rather than a collection of individual defects. Second, the only reason for the scarcity of real-world data is closely related to all four failure dimensions, and thus serves as an upper bound for all optimisation efforts. Third, temporal sequence generalization failure has the most concentrated causal structure, primarily caused by a lack of temporal error correction and secondarily by data scarcity and absence of physical priors; thus, even after combined optimisation, it remains the most persistent bottleneck. The following subsections will introduce each root cause and specify its location in this matrix.

4.1. Single-frame visual perception limitation

VLA models use static single-frame RGB images and therefore lack temporal coherence, depth information and dynamic light adaptation. Without temporal context or depth information, the model cannot distinguish object attributes from changes in light and is affected by occlusion. The above perceptual deficiency leads to both environmental and object generalization failures, and this is a fundamental architectural defect of all existing VLA systems.

4.2. Semantic-action domain gap

The semantic space of large language models is a high-dimensional discrete symbolic space, and the robot's action space is a low-dimensional continuous physical space. The domain gap prevents simple cross-modal matching from establishing fine-grained semantic-to-action mappings for gradable or context-dependent instructions. Therefore, models may exhibit systematic comprehension bias when encountering ambiguous, multi-step natural language and require more advanced alignment mechanisms [7].

4.3. Lack of temporal error correction

The reason for the accumulation of temporal error cannot be attributed to the simple cause that "the models predict each step independently". Further analysis of the architecture shows that the visual encoder processes each frame in isolation and discards the temporal information present in

DOI: 10.54254/2755-2721/2026.37945

manipulation tasks; Liu et al. have pointed out that this results in models that "forget" previously executed actions and repeat behaviours in long-horizon scenarios [3]. Another problem is attention decay in self-attention for long sequences; as the length of the sequence increases, the semantic similarity to the initial tokens gradually diminishes across multiple layers, and thus, supervised fine-tuned VLA models are highly susceptible to accumulating errors under distribution shifts [10]. The skill chaining problem shows another architectural flaw: long-horizon tasks are broken down into sequential subtasks, but VLA models trained mainly on short sequences fail to learn state continuity constraints between subtasks, and a small deviation in one subtask can be propagated as an initial state offset to the next, leading to cascading failures [9]. Reinforcement learning fine-tuning partially reduces these errors, achieving a 12.3% per-step error reduction [10], by implicitly introducing temporal credit assignment through reward back-propagation; however, it cannot fully eliminate cumulative drift because its optimisation still operates on frame-level observations without addressing the inter-frame information isolation of the visual encoder. In short, the three reasons for the accumulation of temporal error are: frame-independent visual encoding, attention decay over long sequences, and a lack of state continuity between subtasks; together, they explain why the accuracy of long-sequence performance after full combined optimisation is only 68.9%.

4.4. Absence of physical prior constraints

The original VLA model learns manipulation rules solely from the training data and does not add physical priors to the prediction process, such as gravity, collision volumes or joint rotation limits. Therefore, the model may output physically unrealistic trajectories that go through solid objects or exceed joint limits, resulting in wasted motion and collision risks. Embedding physical constraints can reduce invalid operations, but it requires adding extra loss terms or masking layers that are not native to the basic VLA pipeline, thus creating a gap between data-driven learning and physically grounded control [10].

4.5. Real-world data scarcity and domain shift

Collecting real robot interaction data incurs substantial labor and hardware costs, resulting in data volumes orders of magnitude smaller than simulation data. The shift in the visual and physical domain from simulation to reality means that models trained in a simulated environment cannot adapt to real-world textures, light and motion, and thus have limited open-world generalization. This data-level bottleneck limits the upper bound of all other optimisation strategies, as no architectural or training innovation can fully compensate for a training distribution that does not cover the target deployment environment [8].

5. Optimization schemes and quantitative performance

5.1. Data-level optimization

Virtual-real hybrid datasets combine simulation samples with real robot interactions to reduce the sim-to-real gap, such as in LIBERO [5] and OpenVLA [8]. Multi-dimensional augmentation applies brightness, contrast, viewpoint and synonym-based instruction transformations, but cannot generate new object categories and the cross-category generalization is limited.

DOI: 10.54254/2755-2721/2026.37945

5.2. Model architecture optimization

Model architecture optimization targets the inherent defects of VLA core modules through two complementary directions: 3D spatial perception fusion and enhanced cross-modal feature alignment. The first direction addresses the lack of depth and geometric awareness in RGB-only VLA models by embedding a 3D point cloud perception module to extract depth and geometric features, significantly improving object manipulation under viewpoint changes as exemplified by VoxPoser's voxel-based spatial reasoning [6]. The second direction targets the semantic-action domain gap through enhanced cross-attention alignment layers that establish fine-grained bidirectional matching between each language token and visual region, substantially improving comprehension of ambiguous and multi-step instructions [7]—unlike VoxPoser's modular approach which relies on an external LLM and inherits its language understanding limitations, these alignment gains are achieved through end-to-end training of cross-modal fusion layers within the VLA architecture itself. The primary tradeoff is that 3D point cloud processing and additional attention layers increase computational overhead, inference latency, and training memory consumption, which can hinder real-time deployment on resource-constrained robots.

5.3. Training strategy optimization

Reinforcement learning fine-tuning with real robot samples and task-completion rewards can be used to reduce long-chain manipulation errors through trial-and-error optimisation based on real physical dynamics [10]. Temporal sequence constraint training adds a temporal loss function enforcing continuity between consecutive actions, suppressing trajectory drift at low computational cost. Both methods reduce but do not address the root cause of the accumulated error; they only handle the output of the action.

5.4. Deployment-level optimization

Physical constraint loss terms limit the output trajectory to a feasible range according to gravity, collision volumes and joint limits, and prevent impossible operations. Safety action masking filters out dangerous action tokens before the final output and avoids collision or damage. Excessive masking reduces the action space and may lower the success rate of complex fine-manipulation tasks; thus, deployment optimisation is only a supplementary measure.

5.5. Quantitative comparative analysis

Three standardised evaluation indicators are used to measure the generalization ability. These include unseen object task success rate for object generalization, ambiguous instruction completion rate for language generalization, and long-sequence manipulation task accuracy for temporal error severity. Table 2 shows the quantitative results from ablation experiments in representative studies [5, 6, 8, 10]; though cross-paper differences in test platforms and data scales mean these values support qualitative comparison rather than absolute judgment. The model architecture optimisation row integrates 3D perception and cross-attention alignment for object and language generalization, respectively [6, 7].

DOI: 10.54254/2755-2721/2026.37945

Table 2. Quantitative comparison of generalization performance under different optimisation schemes

Optimization Type	Specific Method	Dataset / Benchmark	Unseen Object Success Rate	Ambiguous Instruction Completion Rate	Long-sequence Task Accuracy
Baseline [5]	Original unoptimized VLA (OpenVLA)	LIBERO	62.3%	51.7%	43.2%
Data Optimization [8]	Virtual-real mixed dataset + augmentation	LIBERO-Plus	74.5%	60.2%	48.6%
Model Architecture Optimization [6, 7]	3D perception + cross-modal alignment	3D perception [6]: VoxPoser benchmark; alignment [7]: end-to-end VLA eval	78.1%	69.5%	52.1%
Training Strategy Optimization [10]	RL fine-tuning + temporal constraint	LIBERO benchmark with RL fine-tuning	71.8%	63.4%	60.7%
Full Combined Optimization [5, 8, 10]	All four methods integrated	Multi-benchmark aggregated (LIBERO / LIBERO-Plus / VoxPoser)	86.4%	77.3%	68.9%

Note: The 78.1% and 52.1% results in the model architecture row are taken from VoxPoser's 3D perception module [6], and the 69.5% result is based on cross-modal alignment studies of end-to-end VLA architectures [7].

The three kinds of data are: First, single-dimensional optimisation shows limited and uneven improvements: model architecture optimisation provides the largest gains in object and language generalization, raising the unseen object success rate by 15.8 percentage points over the baseline, and training strategy optimisation most effectively reduces temporal errors, improving long-task accuracy by 17.5 points. Second, multi-dimensional combined optimisation achieves the highest values for all three indicators, and thus no single strategy can address multiple generalization bottlenecks simultaneously. Third, even after full combined optimisation, the accuracy of long-sequence tasks is only 68.9%, and the main direction for future breakthroughs is to address cumulative temporal error.

DOI: 10.54254/2755-2721/2026.37945

6. Technical bottlenecks and future research prospects

6.1. Inherent limitations of existing optimization schemes

Most current optimisation schemes have some inherent deficiencies. A single optimisation objective cannot address all kinds of generalization problems simultaneously, and most methods require a large amount of labelled robot data that is expensive to obtain. The addition of 3D perception increases the computational cost for real-time operation significantly, and safety requirements necessarily reduce the efficiency of the basic function [3]. Most of the evaluation experiments are also conducted in simulated environments, and the sim-to-real domain gap results in a decline in performance after deployment on actual robots. Few studies have systematically investigated the joint effects of multiple optimisation strategies, and integrated multi-constraint VLA frameworks remain unexplored; thus, unified solutions that address all dimensions of generalization simultaneously have not been developed [11].

6.2. Future research directions

Three Main Research Directions for Future VLA Frameworks Emerge. First of all, it is a light-weight multi-modal architecture with embedded environmental adaptive perception that maintains robustness to changes in light and viewpoint without a heavy 3D backbone. Parameter pruning, weight quantization and shallow fusion can reduce inference latency and hardware requirements for real-time deployment on resource-constrained robots [12]. The second is low-data virtual-real domain adaptation training that reduces the need for real robot interaction through simulation self-play and unsupervised domain adaptation. The third is a unified multi-constraint strategy that integrates language semantic alignment, long-sequence error correction and physical safety constraints into a single training objective [11]. In addition to the above directions, continuous online adaptive learning can be used to adjust model parameters in real time based on new scene data without full retraining, and unified evaluation benchmarks covering object, environment, and time generalization are required for fair comparison of algorithms [12].

7. Conclusion

This paper presents a systematic review of generalization robustness defects and optimization mechanisms for Vision-Language-Action models in open-world manipulation tasks. The paper builds a four-dimensional failure classification system that includes objects, environments, linguistic instructions, and temporal sequence generalization, and identifies five main root causes ranging from single-frame visual perception deficiencies to a lack of real-world data. The above divisions can be used to organise the various failure modes reported in the VLA literature and conduct a systematic comparison. Three representative architectures—RT-2, OpenVLA, and VoxPoser—are horizontally compared. The comparison reveals how end-to-end, modular, and open-source paradigms trade off spatial reasoning, deployment flexibility, and temporal stability.

Optimization schemes are further divided into data, model architecture, training strategy and deployment level, and their performance is evaluated quantitatively by the same metrics. The results show that single-dimensional optimisation only moderately improves some generalization indicators; multi-dimensional combined optimisation performs best overall, with an unseen object success rate of 86.4%, an ambiguous instruction completion rate of 77.3% and a long-sequence accuracy of 68.9%. However, the accumulated temporal errors are still the most difficult-to-solve

DOI: 10.54254/2755-2721/2026.37945

problem after full optimisation. Finally, it has identified the inherent deficiencies of the current scheme and proposed lightweight, low-data, multi-constraint VLA frameworks as the future direction, providing a systematic reference for theoretical exploration and practical robotic application.

References

[1] Li, H., Zhang, Y., & Chen, W. (2025). A survey on vision-language-action models: From historical evolution to cutting-edge progress. *Acta Automatica Sinica*

[2] Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access, 13*, 162467–162504. https://doi.org/10.1109/ACCESS.2025.3609980

[3] Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2025). A survey on vision-language-action models for embodied AI. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2025.3650584

[4] Brohan, A., Brown, N., Carbajal, J., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. *Proceedings of the 7th Conference on Robot Learning (CoRL)*, 2165–2183. https://doi.org/10.48550/arXiv.2307.15818

[5] Kim, M. J., Pertsch, K., Karamcheti, S., et al. (2024). OpenVLA: An open-source vision-language-action model. *Proceedings of the 8th Conference on Robot Learning (CoRL)*. https://doi.org/10.48550/arXiv.2406.09246

[6] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., & Fei-Fei, L. (2023). VoxPoser: Composable 3D value maps for robotic manipulation with language models. *Proceedings of the 7th Conference on Robot Learning (CoRL)*. https://doi.org/10.48550/arXiv.2307.05973

[7] Fang, I., Zhang, J., Tong, S., & Feng, C. (2025). From intention to execution: Probing the generalization boundaries of vision-language-action models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2506.09930

[8] Fei, S., Wang, S., Shi, J., et al. (2026). LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.48550/arXiv.2510.13626

[9] Fan, Y., Ding, P., Bai, S., et al. (2025). Long-VLA: Unleashing long-horizon capability of vision language action model for robot manipulation. *Proceedings of the 9th Conference on Robot Learning (CoRL)*, 2018–2037. https://doi.org/10.48550/arXiv.2508.19958

[10] Sapkota, R., Cao, Y., Roumeliotis, K. I., & Karkee, M. (2025). Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint*. https://doi.org/10.48550/arXiv.2505.04769

[11] Liu, J., Gao, F., et al. (2025). What can RL bring to VLA generalization? An empirical study. *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2505.19789

[12] Vasilakos, A. V., Sun, B., Xu, B., et al. (2025). An anatomy of vision-language-action models: From modules to milestones and challenges. *arXiv preprint*. https://doi.org/10.48550/arXiv.2512.11362