ISSN:2790-1688 Volume-15-(2025)

RT-2: Vision-Language-Action Models for Generalizable Robotic Control: A Comprehensive Review

Austin Zhou

Harrow International School, Shanghai, China

austin1233@126.com

Abstract. Recent advances in large-scale vision-language models (VLMs) have opened new pathways for robotic learning and control. Google DeepMind's Robotics Transformer 2 (RT-2) represents a transformative step by integrating pre-trained internet-scale multimodal models with physical robotic systems. RT-2 reconceptualizes robot actions as a symbolic language, enabling seamless knowledge transfer from web-scale data to low-level manipulation through action tokenization and co-fine-tuning. This review presents a comprehensive analysis of RT-2's architecture, training methodology, and empirical performance, highlighting its significant improvements in zero-shot generalization, emergent reasoning abilities, and real-world deployment capacity. The paper also examines critical limitations such as physical precision, safety concerns, and computational requirements, offering insights into future directions for scalable, embodied artificial intelligence. RT-2 stands at the forefront of general-purpose robotics and sets the foundation for more capable and adaptive human-robot interaction systems.

Keywords: RT-2, Vision-Language Models, Action Tokenization, General-Purpose Robotics, Zero-Shot Learning

1. Introduction

Robotics has long been hindered by the entrenched perception-planning-execution framework, where discrete modules process simplified information sequentially. This fragmented structure inherently limits a robot's ability to adapt to new contexts, navigate dynamic environments, or understand rich semantic commands. Google DeepMind's RT-2 represents a paradigm shift by directly connecting vision-language models with robotic control to form an end-to-end system [1].

The core insight behind RT-2 is treating actions as a form of language. By tokenizing continuous robotic actions into discrete symbolic representations—ones that can be efficiently processed by large language models—RT-2 merges visual perception, linguistic comprehension, and action synthesis into a single transformer-based architectural design [1]. This approach eliminates the need for human-engineered intermediate representations, allowing knowledge from web-scale data (originally used only for digital tasks) to be transferred directly to the physical control of robots.

This review dissects the RT-2 framework from multiple angles: architectural innovations, training strategy, zero-shot generalization capabilities, emergent behaviors, and real-time deployment performance. Furthermore, it critically examines its current limitations—including manipulation precision, safety robustness, and computational scalability—and outlines research trajectories necessary to achieve broadly deployable, general-purpose robots. As robotics transitions from task-specific automation to adaptive agents, RT-2 marks a foundational step toward that future.

2. Technical Underpinnings from VLMs to VLAs

2.1 Vision-Language-Action (VLA) Model Architecture

RT-2's architecture is built on two vision-language models: the large 55-billion-parameter PaLI-X and the smaller 12-billion-parameter PaLM-E [2,3]. Its key conceptual advance is modeling robotic actions as an extension of language modeling, rather than treating them as discrete control commands. This is achieved through action tokenization, a new method for discretizing continuous robotic actions into a symbolic vocabulary.

The 6-DoF end-effector control (3D position (x, y, z), 3D rotation (roll, pitch, yaw), and gripper open/close) is each discretized into 256 unique bins, with each dimension represented as an integer

between 0 and 255. Importantly, the vocabulary for these action tokens is created by repurposing the 256 least-used tokens in the VLM's existing vocabulary, ensuring compatibility with the model's tokenizer [1]. This discretization retains sufficient accuracy for robotic manipulation while integrating seamlessly into the language model's sequence generation framework.

The unified input-output interface embodies this paradigm. Inputs consist of multimodal data: one or more visual observations (typically RGB images) and a natural language instruction structured as a question in standard Visual Question Answering (VQA) format (e.g., "pick up the blue block near the cup"). The output is an autoregressively generated sequence of action tokens [1]. The model's vocabulary is restricted to valid action tokens during inference, preventing nonsensical outputs. Both visual and text inputs pass through a single transformer backbone. Images are first encoded into a sequence of patches using a Vision Transformer (ViT), followed by feature extraction [1]. The tokenized text instruction is then concatenated with these visual embeddings and fed into the transformer decoder, which autoregressively predicts the next action token based on the input context—effectively "translating" the instruction into executable robotic actions within the learned token space. This end-to-end process avoids intermediary representations like object detection bounding boxes or symbolic scene graphs.

2.2 Training Strategy: Co-Fine-Tuning

RT-2's strong generalization capacity stems from its novel co-finetuning strategy, which blends extensive web-scale semantic knowledge with focused robotic control data [4]. The training corpus includes diverse datasets: over 1 billion image-text pairs from the WebLI dataset (covering 109 languages) to learn general visual recognition and cross-lingual understanding; the RT-1 dataset, consisting of 17 months of real-world demonstrations from 13 kitchen robots (over 130,000 episodes) to provide foundational manipulation skills [4]; and auxiliary VLM data, such as standard VQA datasets for question-answering, image captioning datasets for descriptive language generation, and multimodal mappings to improve cross-modal alignment.

A key innovation is balanced batch composition. Simply combining data sources could lead to catastrophic forgetting of web knowledge or incomplete adaptation to robotic skills. RT-2 uses a conservative weighted ratio: 50% robot data and 50% web data per batch for the RT-2-PaLI-X model, and 66% robot data with 34% web data for RT-2-PaLM-E. Dynamic sampling weights adjust the probability of each data source during training to maintain balance. At the same time, a curriculum learning strategy gradually increases the proportion of robot data—allowing the model to build a solid semantic foundation before refining physical control. Robustness to real-world variations (e.g., lighting, viewpoint, occlusion) is enhanced through random cropping, color jittering, rotation, and Gaussian noise applied to visual inputs [1].

2.3 Real-Time Deployment

Deploying multi-billion-parameter models like RT-2-PaLI-X-55B for real-time robotic control requires addressing several engineering challenges. Distributed cloud inference is used, with models running on Google Cloud TPU v4 Pods and employing state-of-the-art model parallelism to distribute computation across multiple accelerators. To meet operational frequencies of 1–3 Hz for the 55B model and 5 Hz for a smaller 5B variant, latency is optimized through model quantization (reducing the numerical precision of weights), optimized attention mechanisms, and caching intermediate activations for recurrent components in the transformer decoder. A multi-robot serving infrastructure allows a single cloud service to efficiently handle requests from multiple robots concurrently, using request batching, elastic resource allocation, and priority scheduling. Bandwidth is reduced by compressing high-resolution images into efficient codecs like JPEG 2000 before cloud upload, with incremental updates minimizing network traffic. Finally, a safety layer filters the model's output token sequences at the action level, restricting them to safe kinematic ranges to avoid physically undesirable or dangerous actions.

3. Experimental Breakthroughs: Generalization & Emergence

3.1 Unprecedented Zero-Shot Generalization

RT-2's most compelling achievement is its zero-shot generalization—performing novel tasks on which it was never trained. Across over 6,000 physical trials, it consistently outperforms existing state-of-the-art models like RT-1 and MOO (Mobile ALOHA + Open X-Embodiment) [3]. Across all tested scenarios (new items, backgrounds, and environments, both complicated and straightforward), RT-2 (including the PaLI-X variant) consistently performed better, frequently by significant margins, particularly in challenging novel background tasks. Also, it performed exceptionally well in the Language-Table simulation.

RT-2 generalized exceptionally well to over 50 entirely novel objects (e.g., new toys, uniquely shaped containers, special-purpose utensils) by using visual-semantic knowledge from web data to infer object affordances (e.g., recognizing a new container could be grasped despite never seeing it before). Environmental transfer was also strong: trained in kitchens, it performed well in offices, correctly parsing commands like "recycle the soda can" by identifying a blue bin as a recycling bin—demonstrating spatial understanding and task generalization skills learned from web semantics. Additionally, RT-2 executed out-of-language commands, showing the transfer of multilingual understanding from WebLI to physical control, significantly improving accessibility for diverse users.

3.2 Emergent Semantic Capabilities

One of RT-2's most remarkable and unexpected results is the emergence of "new" abilities—skills never explicitly trained but observed in the model, stemming from the transfer of abstract knowledge from web-scale pretraining. These include symbolic reasoning: RT-2 interpreted complex symbolic signs, successfully executing commands like "Move the apple to the location marked 3" by reading alphanumerics on markers, or "Push the coke can to the red heart mark" by locating and acting on semantic icons—demonstrating an understanding that symbols guide action [1].

Relational understanding was another strength; the model learned spatial and semantic relationships between objects, such as proximity ("Pick up the cup next to the plate"), comparative attributes ("Take the smallest fruit"), and semantic similarity ("Put the banana in the yellow bowl"). This allowed it to execute complex instructions requiring simultaneous recognition of multiple objects and their properties, such as "Move the apple to the cup that matches the apple's color."

RT-2 also showed emergent social interaction, recognizing simple human traits in camera feeds (e.g., glasses, hair color, poses indicating "tiredness" or "pointing") to fulfill contextually rich commands like "Hand the bottle of water to the person with glasses" or "Give the energy drink to the tired-looking person." Though basic, this represents a significant step toward humanoid robots operating in human-centric environments.

3.3 Chain-of-Thought (CoT) Reasoning

By enhancing the instruction format to include a "Plan" command, RT-2 achieved a higher level of multi-step reasoning than RT-1, enabling more complex action planning and emergent higher-order behavior. Physical reasoning allowed it to select and use tools: for example, when told to "Hammer the nail into the board," it would first find and pick up a rock to use as a hammer; for "Reach the book on the high shelf," it might locate and position a sturdy box as a step stool.

Contextual decision-making was evident: given "Prepare a snack for the tired person," it consistently chose an energy drink over an apple; for "Prepare a healthy snack," it selected fruit. RT-2 could also parse commands to organize objects, such as "Group the objects by color," resulting in color-coordinated clusters. Complex multi-step actions (e.g., "Make coffee: find cup and coffee machine, put the cup, press the button") were automatically decomposed into steps in the "Plan" output. There was even evidence of hypothetical simulation, with RT-2 briefly "considering" alternative actions internally before generating the final sequence—suggesting rudimentary planning [1].

4. Critical Limitations & Challenges

4.1 Fundamental Constraints

Despite its success, RT-2 has inherent limitations restricting immediate application. An action skill bottleneck exists: its manipulation abilities are fundamentally limited by the RT-1 demonstration data, which consists of simple pick, place, and push actions in tabletop kitchen setups. It lacks real-world dexterity, such as advanced tool use beyond simple substitution, fine motor skills like folding cloth or twisting knobs, and new motion types not in the training distribution [4]. Due to discretization limits and coarse motor commands, it cannot perform precision tasks requiring sub-millimeter accuracy, such as those suited for the PR2 robot.

Dynamic interaction deficits are another issue: RT-2 struggles with even basic physical objects, with failure rates exceeding 80% for non-rigid items (e.g., bananas), rolling objects (e.g., pens), flexible wires, and fragile items (e.g., eggs) [1]. It cannot execute complex chain reactions involving multiple interacting objects (e.g., domino effects) and shows minimal reasoning about material properties (e.g., differentiating rubber and rigid metal), failing at tasks involving fluids (pouring water) or granular materials (scooping rice)—revealing a lack of intuitive physics modeling [3].

Finally, RT-2's operational scale is impractical for widespread deployment: the RT-2-PaLI-X-55B architecture requires a cluster of 8 NVIDIA A100 GPUs or access to Google Cloud TPU v4 Pods. Its inference latencies (300–1000 ms, equivalent to 1–3 Hz) are too slow for tasks like catching falling objects or agile locomotion, and its power consumption (~500 W) is infeasible for battery-operated mobile robots [1]. This creates a critical reliance on the cloud, with associated issues like network latency and availability.

4.2 Safety & Robustness Concerns

Also, safety and reliability concerns are to be handled for practical deployment outside of labs. Action precision errors are inherent in discretization, giving ±0.5 cm positional and ±3° rotational errors—acceptable for coarse tasks but not for insertion or working near obstacles [5]. RT-2 has no internal state model and can only react to each state's task. It cannot keep track of something such as whether a drawer is open between several steps, producing incoherent sequences.

Adversarial susceptibility is a second vulnerability: system performance deteriorates dramatically in challenging situations with sudden illumination changes, partial object occlusions, visual noise, or adversarial patterns crafted to fool vision models. The one is no longer outfitted with any integrative safety concept that counteracts crashes, obstacles, identified forces, or emergent safety braking—all this needs to be applied as safety shells from the outside. AOnet, furthermore, does not provide uncertainty readouts; it cannot express uncertainty about whether a particular action should be taken, and one's sure policies might be incorrect [5]. Finally, RT-2 cannot reason about failures or unforeseen environmental changes when executing tasks, which restricts failure recovery.

4.3 Failure Case Analysis

Further failures of the Language-Table benchmark were investigated in more detail. Performance for dynamic object manipulation fell short of the 20% success rate for non-rigid objects such as bananas or pens because it was difficult to anticipate complex physical interactions. In addition, poor estimates of center-of-mass effects and friction coefficients across materials were also found.

Part-level interaction scores (grabbing a handle, unscrewing a bottle) achieved a 32% success rate due to poor spatial resolution necessary for visual feature extraction, difficulty in interpreting the kinematics of articulated objects, and occlusion of body parts during manipulation [6]. For tasks longer than three steps in a row, success rates were more than 40% lower than for single-step tasks, with cumulative errors, no working memory for intermediate goals or states, and no backtracking or replanning following partial failures all contributing to this loss.

RT-2 also has difficulty with vague tasks that rely on common sense (e.g., "clean the table," "prepare the table for breakfast," "find something unhealthy," etc.). It knows nothing about human

organizational habits, cannot guess the user's unspoken preferences, and has no way to clarify queries further.

5. Future Directions: Scaling Embodied Intelligence

5.1 Architectural Innovations

Several architectural improvements are needed to overcome current limitations. Lightweight VLA models must be developed for real-time on-robot deployment, using strategies like knowledge distillation (training RT-2-Tiny models for specific tasks with >80% parameter reduction), aggressive quantization (FP16, INT8, or binary), structured/unstructured pruning, and sparse activation patterns. Hybrid architectures combining transformer-based semantic power with traditional MPC or behavior trees for low-level stability are also valuable [7].

Strong multi-modal fusion is essential: physical interaction requires more than vision and language, incorporating tactile sensor data for fine manipulation, audio inputs and outputs for better human-robot interaction, proprioceptive data and force feedback loops for compliant, safe contact, depth sensing for enhanced 3D spatial awareness, and even rudimentary olfactory sensing for material identification [7].

Memory and state tracking mechanisms are necessary for long-horizon tasks, including learnable storage-based attention with dual read heads controlling memory access, slot-based attention (for tracking objects), predictive world models (in neural or symbolic forms) to simulate action outcomes, Bayesian filtering for state updates during manipulation, and specialized modules for generating failure recovery policies.

5.2 Data & Training Paradigms

Scaling embodied intelligence requires radical data strategies. Cross-platform and cross-embodiment datasets, such as the RT-X initiative (collecting millions of demonstrations across 22 robot morphologies, including bipedal, quadrupedal, and multi-gripper arms), are essential [8]. The Open X-Embodiment dataset (with 150,000+ tasks) has shown that such data can improve cross-robot generalization by 50% [9]. Augmented sim-to-real transfer methods and failure-focused training benchmarks are also key.

Learning from human videos is another direction; large-scale egocentric video datasets (e.g., Ego4D, YouTube) offer insights into complex unstructured skills [10]. Challenges include cross-modal alignment (translating human actions to robot-friendly motor commands, known as "Learning from Observation") and enabling robots to acquire skills through autonomous, self-supervised environmental interaction. Cross-body imitation learning aims to transfer human motions to diverse robot bodies.

Finally, self-evolution and constant learning are critical: beyond one-time training, robots need lifelong adaptation, using online reinforcement learning with real-world success/failure signals to refine skills continuously [11]. This includes self-supervised exploration for open-loop environmental interaction and skill acquisition, automatic curriculum learning to adjust task difficulty dynamically, adversarial training to enhance robustness to perturbations, and failure-driven learning to focus data collection on challenging edge cases.

5.3 Application Areas/Societal Impact

RT-2’s paradigm enables transformative applications with unique challenges and societal implications. In industrial manufacturing, adaptive assembly lines could process variant products, enable flexible warehouse material handling, and perform semantic-level visual quality inspection—potentially increasing productivity by 30% and enabling mass customization. Challenges include achieving year-round millimeter-level accuracy, embedding certified functional safety systems, meeting challenging real-time control loop requirements, and ensuring explainable decision-making for process safety [12].

Home and assistive robotics could revolutionize elder care (e.g., medication reminders, fall detection, companionship), reduce domestic labor (e.g., cleaning, organization, meal preparation), and adapt to individual user needs. However, reliable operation in unstructured, dynamic home environments; absolute safety during close human-robot interaction; adherence to subtle social norms; and user privacy protection are critical challenges, along with improving long-term reliability and emotional intelligence.

In health and rehabilitation, applications might include assisting surgeons with complex operations (via semantic understanding), recommending physical therapy routines, enabling 24/7 patient monitoring, and aiding disabled individuals. Challenges include meeting stringent safety and sterility requirements, navigating complex ethical reviews and FDA/CE regulatory approvals, acquiring deep clinical expertise, ensuring clinician oversight, and maintaining detailed audit trails [13].

6. Conclusion

RT-2 represents a seminal, empirically validated advance in robotic learning, clearly demonstrating that large vision-language models pretrained on vast internet-scale datasets can generate executable physical control commands by framing robotics as a language-domain sequence modeling problem. Its key technical contributions—action tokenization and semantically grounded co-finetuning—achieve what modular systems could not: effective transfer of abstract knowledge (visual concepts, multilingual semantics, spatial reasoning, and contextual understanding) from the digital web to physical robotic interaction [13].

Yet, significant hurdles remain before general-purpose robots are realized: the enormous computational demands of modern VLAs, difficulties with dynamic physical interactions, ongoing safety concerns, and long-horizon planning issues. Addressing these requires aggressive research into efficient, multimodal architectures, cross-embodiment and human-video-based data innovations, and reliable continual learning [8-10]. As the technology matures, researchers must extend their focus beyond just technical capabilities to bring in advances in onboard intelligence, deep learning, unsupervised learning, computer-human interaction, AI ethics, and safety protocols for learned systems and policies with social or economic context.

References

[1] Brohan, Anthony, et al. "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control." arXiv preprint arXiv:2307.15818, 2023.

[2] Chen, Xi, et al. "PaLI-X: On Scaling Up a Multilingual Vision-Language Model." arXiv preprint arXiv:2305.18565, 2023.

[3] Driess, Danny, et al. "PaLM-E: An Embodied Multimodal Language Model." arXiv preprint arXiv:2303.03378, 2023

[4] Lynch, Corey, et al. "Interactive Language: Talking to Robots in Real Time." arXiv preprint arXiv:2210.06407, 2022

[5] Stone, Austin, et al. "Open-World Object Manipulation Using Pre-trained Vision-Language Models." arXiv preprint arXiv:2303.00905, 2023

[6] Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. "CLIPort: What and Where Pathways for Robotic Manipulation." Proceedings of the 5th Conference on Robot Learning, 2022.

[7] Dasari, Sudeep, et al. "RoboNet: Large-Scale Multi-Robot Learning." Proceedings of the 3rd Conference on Robot Learning, 2019

[8] Grauman, Kristen, et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18995-19012

[9] Open X-Embodiment Collaboration. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models." arXiv preprint arXiv:2310.08864, 2023

[10] RT-X Team. "Scaling Robot Learning with Cross-Embodiment Datasets." Google DeepMind Technical Report, 2023.

[11] Levine, Sergey, et al. "Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection." The International Journal of Robotics Research, vol. 37, no. 4-5, 2018, pp. 421-436.

[12] Andrychowicz, Marcin, et al. "Learning Dexterous In-Hand Manipulation." The International Journal of Robotics Research, vol. 39, no. 1, 2020, pp. 3-20.

[13] Gupta, Abhishek, et al. "Unlocking the Potential of Simulators for Deep Robotic Learning." Annual Review of Control, Robotics, and Autonomous Systems, vol. 6, 2023, pp. 177-199.