CVPRW FMEA
#42

EvoWorld: A World-Model-Centric Framework for Continuous Self-Evolution of Modular Embodied Skills

Anonymous CVPR Workshop FMEA submission

Paper ID 42

Abstract

Current progress in physical intelligence largely relies on scaling monolithic Vision-Language-Action (VLA) models, yet real-world policy data remain fragmented across scenes and tasks. This mismatch limits transfer, exacerbates catastrophic forgetting, and impedes continual improvement. A modular design that shares dynamics while specializing skills is therefore a promising paradigm. We introduce ***EvoWorld*** *(****EvoW****), a world-model-centric framework for skill orchestration and iterative self-evolution. In EvoW, VLAs form an expandable library of pluggable experts. A high-level router selects experts conditioned on scene and task, while an action-conditioned video world model serves as the cognitive core for rollout-based planning. The world model provides counterfactual rollouts to score candidate experts, while selected experts execute in the grounded scene to generate trajectories for verification. A vision-language evaluator delivers semantic scoring and diagnostic tags, enabling targeted updates to the world model, router memory, or specific experts rather than global retraining. This closes an automated loop that jointly improves grounding, routing, and skill refinement without manual task engineering. Experiments show that EvoW enables automated task-to-policy synthesis with competitive success rates and iterative improvement trends over iterations, while producing valid and diverse trajectories that support evaluation and skill refinement.*

1. Introduction

Physical intelligence is increasingly required to follow open-ended user intent rather than a fixed list of benchmark tasks [1, 4, 10]. To meet this demand, the field has largely trended toward scaling monolithic Vision-Language-Action (VLA) models and diffusion-style policies in order to capture a broad spectrum of robotic behaviors [5, 7, 25]. Large-scale datasets such as DROID and Open X-Embodiment have further fueled this progress by providing unprecedented task and visual coverage [8, 24]. Despite this, a potential misalignment between universal physical dynamics and highly task-specific execution strategies has been partially overlooked in the pursuit of monolithic scaling, leaving the challenge of pervasive real-world data silos somewhat under-addressed.

Current generalist control approaches rely predominantly on training large VLA models or flow-based policies on heterogeneous data to achieve broad manipulation coverage [3–5, 7, 25, 28], treating diverse robotic experiences as tokens in a shared backbone and assuming that a single, sufficiently large network generalizes across embodiments via a shared latent space. In practice, however, monolithic methods face limitations under fragmented scene distributions and long-horizon compounding errors [2, 26]: robotic data remain distributed across labs and protocols [8, 24], making continual updates costly and vulnerable to forgetting; and without explicit feasibility verification, these systems drift semantically in open-world settings [30, 44].

We identify the root of these challenges in the lack of decoupling between universal physical dynamics and task-specific execution strategies [15, 18, 38]. The laws of physics are largely invariant across scenarios, while the "skills" required for manipulation are diverse and highly context-dependent [3, 25, 28]. Current paradigms lack an explicit mapping that routes task semantics to a suitable specialist, forcing a single network to resolve competing gradients from disparate domains [40, 41]. We argue that a more rational architecture treats the world model as a universal cognitive foundation while maintaining policies as modular, pluggable skills updated independently [18, 38].

EvoWorld (EvoW) is proposed as a world-model-centric framework for task-to-policy automation that enables continuous skill evolution. Unlike monolithic approaches [3, 20, 25, 28], EvoW separates shared physical grounding from task-level execution and treats VLAs as modular components in an expandable library. The framework operates through a principled pipeline: (i) anchoring the task in a high-fidelity physical scene via semantic retrieval, (ii) performing counterfactual simulation using an

CVPRW FMEA
#42

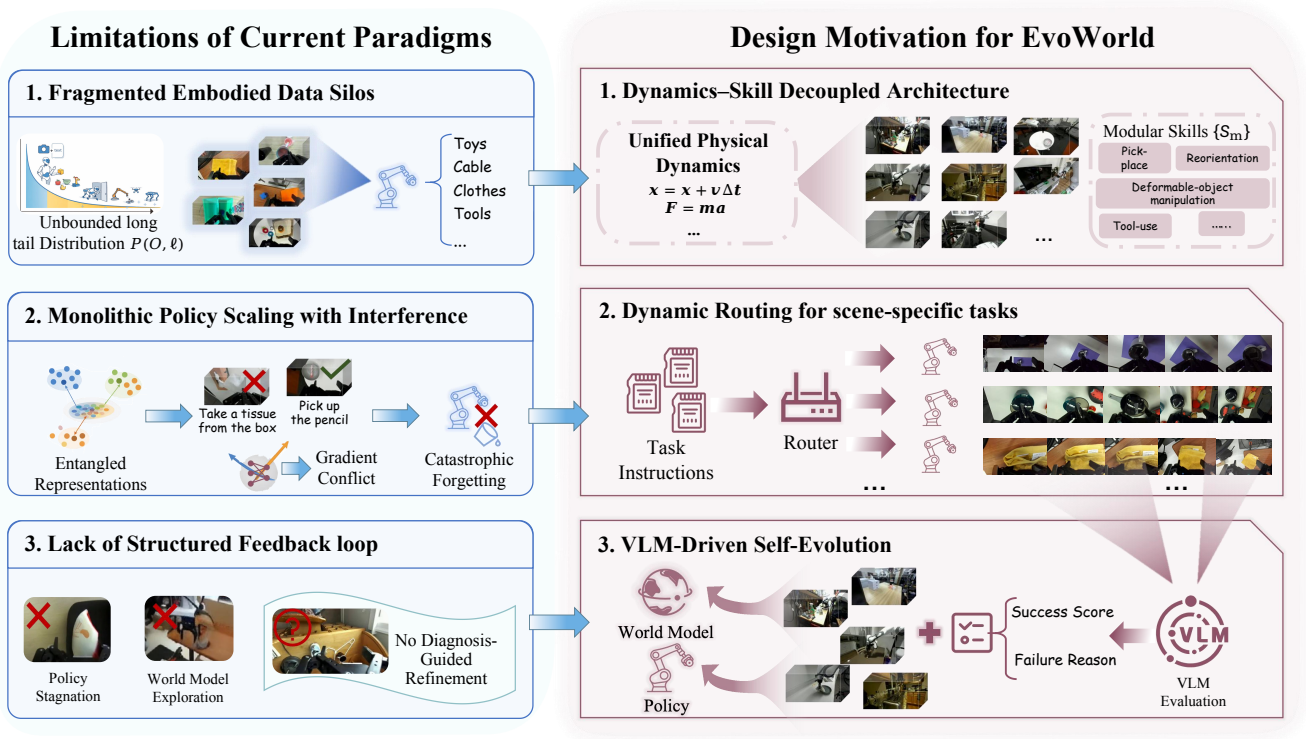

Figure 1. **Motivation and system overview of EvoW. Left:** current embodied-learning pipelines face three structural bottlenecks: (1) fragmented embodied data silos with long-tail task/scene distributions, (2) interference and gradient conflict in monolithic policy scaling that induce catastrophic forgetting, and (3) the lack of a diagnosis-guided feedback loop for iterative refinement. **Right:** the proposed framework resolves these bottlenecks via three coupled components: (1) a dynamics–skill decoupled architecture that separates reusable physical dynamics from modular task-specialized skills, (2) scene-aware dynamic routing that maps task instructions and grounded observations to the most suitable expert policy, and (3) a VLM-driven self-evolution loop that scores outcomes, identifies failure causes, and feeds diagnostic signals back to world-model adaptation and skill updates.

action-conditioned video world model [30, 44], (iii) routing the task to the optimal VLA expert through a scene-aware scheduler, and (iv) driving a closed-loop self-evolution cycle via Vision-Language Model (VLM) diagnosis. This design allows the system to autonomously refine the world model and retrain specific skills without re-initializing the entire intelligence substrate.

Experiments demonstrate that EvoW outperforms monolithic baselines [3, 25, 28] in success rates, particularly across fragmented data silos, and achieves compounding improvements over iterations through VLM-based failure diagnosis and expert routing. We summarize our key contributions as follows:

- **Paradigm:** we propose a principled world-model-centric formulation that disentangles universal physical dynamics from task-specialized skills, shifting the focus from policy scaling to world-model-driven orchestration.
- **Method:** we implement the EvoW framework, which operationalizes this decoupling through retrieval-grounded scene construction and a VLM-driven diagnostic loop that enables autonomous, closed-loop self-evolution.
- **Validation:** extensive experiments across data silos verify the necessity of expert routing and self-evolution, demonstrating that EvoW achieves compounding performance gains without the overhead of full-system retraining.

2. Related Work

Scaling policies and data. Generalist VLA policies such as RT-1, RT-2, and OpenVLA demonstrate strong instruction-following behavior through scale [4, 5, 25], and industrial foundation efforts (e.g., GR00T N1) push this direction with larger model and data pipelines [28]. In parallel, large real-robot datasets including Open X-Embodiment, DROID, and RH20T broaden scene and task coverage [8, 11, 24]. However, most pipelines remain monolithic, where updates are largely global and specialist selection at deployment is weak. Under fragmented data silos, this often leads to interference and higher continual-update cost.

Manipulation policies and structured primitives. Beyond model/data scaling, manipulation research focuses on

CVPRW FMEA #42

control structure, including end-to-end visuomotor learning, reinforcement learning, and imitation learning [23, 27, 35]. Structured formulations such as Transporter Networks and CLIPort improve compositionality and sample efficiency in multi-step tasks via stronger inductive bias for spatial reasoning and action decomposition [34, 42]. Recent dexterity-oriented works further extend policy capacity under richer embodiment and contact settings [43, 45]. Still, these methods are typically optimized within fixed policy-training pipelines and lack a diagnosis-guided, module-specific refinement loop coordinating grounding, routing, and world-model updates.

Grounding and memory for embodied reuse. Grounding and memory-based approaches improve generalization by reusing prior trajectories, grounded plans, and contextual cues [33]. Their core strength is efficient reuse under recurring task patterns, yet in many settings grounding is not tightly coupled with verifiable rollout generation and failure attribution. As a result, they are less suited to autonomous task-to-policy iteration when scenes and task semantics shift together in open-ended real-robot settings.

World models for long-horizon control. World models support imagination-based rollouts for planning and policy improvement [15–17]. Recent latent-action and video-prediction models improve rollout horizon and perceptual realism [12, 13, 39], and other studies investigate tighter coupling between world models and control policies [6, 21, 22]. A core challenge remains long-horizon robustness: compounding prediction errors and semantic drift accumulate over time [2, 26].

Automated task generation and our position. RoboGen shows that closed-loop task generation can scale quickly in simulation [37]. For real-robot deployment, additional constraints are necessary: real-scene grounding, expert routing at deployment, and diagnosis-guided correction. EvoW addresses this gap by combining grounded initialization, task/scene-aware routing, and targeted diagnosis-guided updates in a world-model-centric loop, with explicit decoupling between shared dynamics and task-specialized skills.

3. Method

In this section we delineate the EvoW framework across three strategic dimensions: a formal problem formulation, the modular architectural design, and a diagnosis-driven closed-loop optimization. Our approach emphasizes how Contextual Scene Anchoring, Strategic Routing, and Action-conditioned World-model Rollouts are synergistically coupled to facilitate continuous, autonomous task-to-policy evolution.

3.1. Notations and Problem Setup

We formalize open-ended task-to-policy synthesis as a language-conditioned decision process. Let ℓ denote a natural language task instruction (e.g., "pick up the green object and put it in the bowl"). At each timestep t, o_t represents the visual observation and a_t the executed action. An interaction episode of horizon H is a trajectory $\tau = (o_0, a_0, \ldots, o_H)$. To address the limitations of monolithic scaling in fragmented environments, EvoW decomposes this process into universal dynamics and specialized control:

A universal video world model (p_θ): acting as the system's cognitive engine, this model captures transferable environmental dynamics. The action-conditioned latent transition is

$$z_{t+1} \sim p_\theta(z_{t+1} \mid z_t, a_t), \qquad o_{t+1} = D(z_{t+1}), \tag{1}$$

where E encodes observations into a latent space $z_t = E(o_t)$, and D decodes predicted states back into the observation manifold.

A modular skill library (Π) aggregates task-specific execution strategies as a collection of M specialized VLA experts:

$$\Pi = \{\pi_m\}_{m=1}^{M}, \qquad a_{t:t+H} \sim \pi_m(\cdot \mid o_t, \ell), \tag{2}$$

where a high-level Strategic Router q maps the grounded scene and instruction to the optimal expert π_{m^*} based on the experience memory $\mathcal{M}$:

$$m^* = \arg\max_m q(m \mid o_0, \ell, \mathcal{M}). \tag{3}$$

This world-model-centric decomposition allows physical dynamics to be shared as a universal prior, while policy capacity is allocated through modular specialization and MLLM-driven routing. As illustrated in Figure 2, this architectural decoupling lets EvoW bypass real-world data silos and achieve continuous self-improvement through an introspective diagnostic loop.

3.2. Contextual Scene Anchoring

To bridge the gap between abstract linguistic intent and executable physical behavior, EvoW adopts a Contextual Scene Anchoring paradigm whose objective is to initialize a semantically aligned real-world observation, thereby constraining the subsequent simulation to a physically plausible manifold rather than generating outcomes in a vacuum.

We define an offline experience manifold $\mathcal{D} = \{(o_i, \xi_i)\}_{i=1}^{N}$, where each entry represents a high-fidelity snapshot of real-robot interaction with observation o_i and auxiliary context ξ_i. Given an open-vocabulary task specification ℓ, we project the language instruction and visual candidates into a shared latent embedding space:

$$e_\ell = \Phi_L(\ell), \qquad e_{o_i} = \Phi_O(o_i), \tag{4}$$

CVPRW FMEA
#42

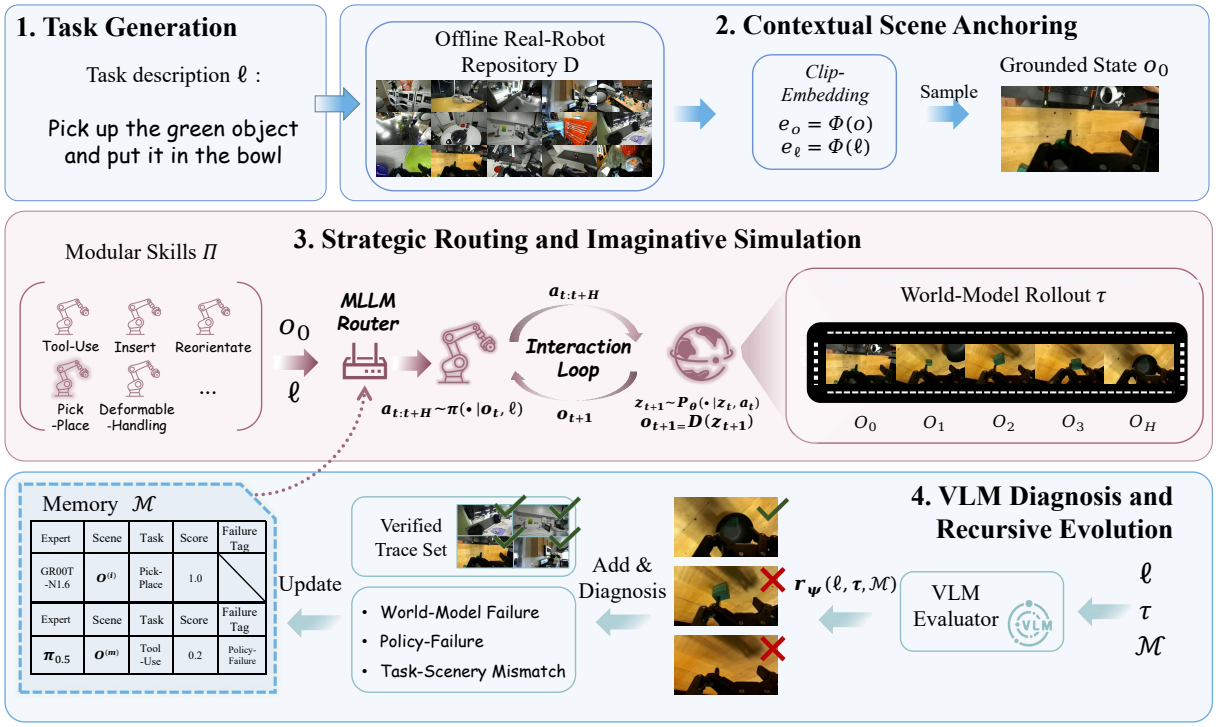

Figure 2. **EvoWorld pipeline overview. (1) Task Generation:** construct candidate task specifications for each round. **(2) Contextual Scene Anchoring:** select a semantically aligned real scene from offline experience. **(3) Strategic Routing and Imaginative Simulation:** score candidates via world-model imagination, route to a suitable VLA expert, and execute the selected expert. **(4) VLM Diagnosis and Recursive Evolution:** use success score and failure type to trigger targeted module updates rather than global retraining.

with Φ_L, Φ_O pre-trained dual encoders optimized for cross-modal alignment. To ensure diverse yet relevant initialization, we ground ℓ by sampling an initial state o_0 from a semantic-similarity-based softmax:

$$o_0 \sim p(i \mid \ell) = \frac{\exp(\langle e_\ell, e_{o_i} \rangle / \tau_g)}{\sum_{j=1}^{N} \exp(\langle e_\ell, e_{o_j} \rangle / \tau_g)}, \tag{5}$$

where τ_g controls the exploration–exploitation trade-off during grounding.

Once o_0 is anchored, the action-conditioned video world model serves as the primary cognitive engine for counterfactual simulation. Encoding o_0 into $z_0 = E(o_0)$, the system generates a temporal rollout via recursive latent transitions $z_{t+1} \sim p_\theta(z_{t+1} \mid z_t, a_t)$ and decoding $o_{t+1} = D(z_{t+1})$. This factorization explicitly disentangles scene grounding from dynamics prediction, mitigating the compounding errors and semantic drift that typically plague long-horizon generative models.

3.3. Strategic Routing and Imaginative Simulation

Given the fragmented nature of real-world robotic data, a single monolithic policy often suffers from catastrophic forgetting and performance degradation across diverse embodiments. To circumvent this, EvoW treats each specialized VLA as a modular expert π_m within a library $\{\pi_m\}_{m=1}^{M}$, where each expert is optimized for specific task families or physical constraints. The core challenge then shifts to Strategic Routing: mapping an open-vocabulary instruction ℓ and a grounded observation o_0 to the most compatible expert.

We implement the strategic router via a Multi-modal Large Language Model (MLLM) that leverages its semantic knowledge to evaluate the compatibility between task requirements and the functional descriptions of VLA experts. The router is conditioned on a long-term Experience Memory:

$$\pi_{m^*} = q(\ell, o_0, \mathcal{M}), \quad \mathcal{M} = \{(\ell^i, o_0^i, \pi_m^i, y^i, \kappa^i)\}_{i=1}^{N}, \tag{6}$$

where each entry captures task ℓ, initial state o_0, selected expert π_m, performance score y, and diagnostic failure type κ.

The routing process operates in a dual-mode logic of exploration and exploitation. **Exploration:** for unfamiliar task–scene pairs with low semantic density in $\mathcal{M}$, the router encourages diversity by sampling from a broader categorical distribution to discover new "scene–expert" correspondences. **Exploitation:** for recurrent contexts, the MLLM synthesizes historical success/failure cues from $\mathcal{M}$

CVPRW FMEA #42

Algorithm 1 EvoW closed-loop routing and self-evolution

Require: Task set $\mathcal{L}$, manifold $\mathcal{D}$, expert library Π, world model p_θ, router q, evaluator Ψ, memory $\mathcal{M}$, threshold δ.
Ensure: Verified trajectory set $\widehat{\mathcal{T}}$, updated Π, refined $\mathcal{M}$.

$\widehat{\mathcal{T}} \leftarrow \emptyset$
for each round **do**
  (S1) sample task ℓ
  (S2) ground o_0 from $\mathcal{D}$ conditioned on ℓ
  (S3) select $\pi_{m^*} \leftarrow q(\cdot \mid \ell, o_0, \mathcal{M})$; rollout-score candidates; execute π_{m^*} to obtain τ
  (S4) $(r_\psi, \kappa) \leftarrow \Psi(\tau, \ell, \mathcal{M})$; update $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\ell, o_0, m^*, r_\psi, \kappa)\}$
  if $r_\psi \geq \delta$ **then**
    $\widehat{\mathcal{T}} \leftarrow \widehat{\mathcal{T}} \cup \{(\tau, \ell, o_0, m^*, r_\psi)\}$
  else if $\kappa = \text{TSM}$ **then**
    update grounding module
  else if $\kappa = \text{PF}$ **then**
    update routing and/or selected expert
  else
    update world model
  end if
end for
return $\widehat{\mathcal{T}}, \mathcal{M}$

into memory-weighted logits, biasing selection toward experts with proven reliability in similar manifolds. To further solidify the decision, for each candidate expert π_m identified by the MLLM, the world model generates an imagined rollout τ_{π_m}. The router computes an imaginative-rollout score that reflects predicted task-consistency, and the final expert π_{m^*} maximizes the joint of semantic compatibility and simulated success, ensuring execution is grounded in both high-level intent and low-level physical feasibility.

3.4. VLM Diagnosis and Recursive Evolution

To achieve continuous self-improvement while avoiding the cost and forgetting of global retraining, EvoW implements an introspective diagnosis mechanism leveraging the cross-modal reasoning of VLMs. This stage performs fine-grained credit assignment, pinpointing the bottleneck within the decoupled architecture. Upon executing a routed expert π_m that yields trajectory τ, the VLM evaluator Ψ contrasts the actual outcome with the task specification ℓ and the initial physical grounding:

$$(r_\psi, \kappa) = \Psi(\tau, \ell, \mathcal{M}), \quad r_\psi \in [0, 1], \quad \kappa \in \{\text{TSM}, \text{PF}, \text{WMF}\}, \tag{7}$$

where r_ψ is a continuous success metric and κ a categorical failure attribution: Task–Scene Mismatch (TSM) flags semantic-grounding errors; Policy Failure (PF) flags suboptimal control from the VLA; World-Model Failure (WMF) flags divergence between predicted dynamics and physical reality. EvoW then orchestrates targeted updates above a reliability threshold δ:

Verified experience accumulation ($r_\psi \geq \delta$): high-confidence trajectories are integrated into the verified trace set as high-fidelity demonstrations for self-imitation and router calibration.

Semantic grounding refinement ($\kappa = \text{TSM}$): the system updates router memory $\mathcal{M}$ to refine the semantic-to-scene grounding manifold and prevent recurrence of analogous grounding errors.

Modular policy plasticity ($\kappa = \text{PF}$): autonomous retraining is triggered for the specific VLA expert π_{m^*} using the diagnostic feedback as a corrective signal.

Dynamics grounding refinement ($\kappa = \text{WMF}$): when the simulation fails to capture valid physics, world-model parameters are updated using the discrepancy between τ and the physical ground truth.

This targeted evolutionary strategy allows EvoW to bypass the one-size-fits-all training bottleneck and achieve continuous compounding improvements through recursive, module-specific plasticity. The full procedure is summarized in Algorithm 1.

4. Experiments

We evaluate EvoW along three dimensions: **(1) skill-level performance** (single-skill SR and iterative policy improvement); **(2) task diversity and scene validity**; and **(3) generative quality** of imagined trajectories.

4.1. Experimental Setup

To isolate the contributions of our method, we keep VLA backbones, world-model data sources, and evaluation splits fixed across compared variants. We sample tasks from skill families and use grounded initial scenes anchored to a shared offline repository. Across variants, we fix the rollout horizon, task split, scene split, and trajectory budget per task. We first compare base policies and routed variants under identical initialization, then run iterative refinement with matched update budgets per round.

The benchmark covers five representative manipulation families $\mathcal{S}_1$–$\mathcal{S}_5$: $\mathcal{S}_1$ (Pick-and-Place), $\mathcal{S}_2$ (Tool-Use), $\mathcal{S}_3$ (Reorientation), $\mathcal{S}_4$ (Deformable-Object Manipulation), and $\mathcal{S}_5$ (Articulation).

4.2. Skill-Level Performance

We first study whether routing alone enhances policy success under fixed model and data constraints, then analyze cumulative benefits from iterative refinement. Table 1 reports static SR across four VLA backbones and the five skill families. Routing with EvoW consistently improves success rates: averaged over backbones, SR rises from 0.259 to

CVPRW FMEA
#42

Table 1. Policy success rates across five skill families and four VLA backbones. **EvoW*** denotes our routed variant; Gain(↑%) is relative to the corresponding base average.

Backbone	Setting	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	Avg	Gain(↑%)
GR00T-N1.6	Base	0.38	0.24	0.18	0.28	0.20	0.256	–
	EvoW*	**0.48**	**0.50**	**0.46**	**0.42**	**0.48**	**0.468**	**↑82.8%**
PI-0.5	Base	0.32	0.14	0.24	0.18	0.28	0.232	–
	EvoW*	**0.48**	**0.28**	**0.38**	**0.44**	**0.48**	**0.412**	**↑77.6%**
PI-0	Base	0.40	0.22	0.26	0.32	0.22	0.284	–
	EvoW*	**0.62**	**0.38**	**0.40**	**0.40**	**0.50**	**0.460**	**↑62.0%**
PI-0-fast	Base	0.38	0.18	0.26	0.24	0.26	0.264	–
	EvoW*	**0.66**	**0.38**	**0.44**	**0.36**	**0.42**	**0.452**	**↑71.2%**

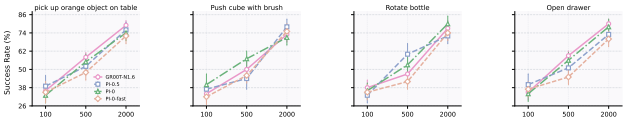

Figure 3. Task-level refinement trends. Iterative diagnosis-driven updates raise SR; variance bars indicate stable gains across runs.

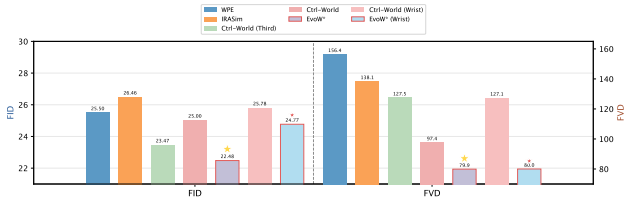

Figure 4. Generative-quality comparison under a view-conditioned setting. EvoW achieves lower FID and FVD than the compared baselines; the larger reduction on FVD is consistent with improved temporal coherence.

0.448 (absolute +18.9 points, relative +73.0%), with per-backbone relative gains spanning +62.0% to +82.8%, indicating the gain is stable rather than concentrated on a single backbone.

Figure 3 reports SR over refinement iterations on representative tasks. SR improves across tasks, indicating that diagnosis-conditioned updates in Stage 4 reduce recurrent failure patterns over time. These trends are consistent with the formulation: expert selection contributes immediate gains under fixed backbones, while diagnosis-conditioned refinement yields additional improvements over successive rounds.

4.3. Task Diversity, Scene Validity, and Generative Quality

Beyond execution success, EvoW must scale task generation while preserving scene validity, and its imagined rollouts must remain visually faithful and temporally consistent.

Diversity vs. executability. Compared with simulation generators including RoboGen [37], RLBench, MetaWorld, ManiSkill2 and GenSim, EvoW achieves the lowest Self-BLEU [29] and CLIP image similarity [31], with competitive ViT similarity [9]. EvoW is weaker on Sentence-BERT similarity [32], which is expected because grounded executability constraints reduce free-form language variation. This trade-off is acceptable: tasks are intended for physically realizable rollouts rather than unconstrained textual diversity.

Generative quality. Against world-model baselines (WorldGym, IRASim, Ctrl-World) [14, 30, 44], we report FID [19] and FVD [36]. Figure 4 shows that EvoW achieves lower FID and FVD than the compared baselines; the larger relative gain on FVD indicates stronger temporal consistency over long-horizon rollouts. Under view-conditioned settings, EvoW remains better than Ctrl-World for both third-view and wrist-view comparisons, consistent with improved rollout stability and better preservation of object identity and contact progression across frames.

5. Conclusion

We presented **EvoW**, a framework that maps open-ended task semantics to verified training trajectories and executable manipulation policies. EvoW grounds each task in task-consistent real scenes, synthesizes scene-consistent episodes with a decoupled world model, and closes the loop via VLM-based routing and verification. This design decouples shared dynamics from task-specialized skills and enables targeted updates through failure diagnosis. Scene grounding and automated VLM feedback are complementary for scaling open-ended manipulation, yielding reliable execution and continual improvement over iterations. Future work includes (i) expanding the anchor repository, (ii) uncertainty-aware long-horizon rollouts and filtering, and (iii) calibrated routing/verification with task-aware critics.

CVPRW FMEA #42

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances. In *CORL*, 2022.

[2] Kavosh Asadi, Dipendra Misra, Seungchan Kim, and Michel L. Littman. Combating the compounding-error problem with a multi-step model, 2019.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CORL*, 2023.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *RSS*, 2023.

[6] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model, 2025.

[7] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023.

[8] Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi,

CVPRW FMEA #42

Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2025.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[10] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *ICML*, 2023.

[11] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot, 2023.

[12] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. In *ICML*, 2025.

[13] Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild, 2026.

[14] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *ICLR*, 2026.

[15] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

[16] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

[17] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *ICLR*, 2021.

[18] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[20] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025.

[21] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models, 2025.

[22] Yuxin Jiang, Yuchao Gu, Ivor W. Tsang, and Mike Zheng Shou. Olaf-world: Orienting latent actions for video world modeling, 2026.

[23] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, 2018.

[24] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia,

CVPRW FMEA
#42

Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. In *RSS*, 2024.

[25] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *CORL*, 2024.

[26] Nathan Lambert, Kristofer Pister, and Roberto Calandra. Investigating compounding prediction errors in learned dynamics models, 2022.

[27] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 17(39):1–40, 2016.

[28] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, 2002.

[30] Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. Worldgym: World model as an environment for policy evaluation, 2025.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[32] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conf. Empirical Methods Natural Lang. Process.*, 2019.

[33] Gabriel Sarch, Yue Wu, Michael J. Tarr, and Katerina Fragkiadaki. Open-ended instructable embodied agents with memory-augmented large language models. In *EMNLP*, 2023.

[34] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *CORL*, 2021.

[35] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation, 2018.

[36] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2018.

[37] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. In *ICML*, 2024.

[38] Yucen Wang, Fengming Zhang, De-Chuan Zhan, Li Zhao, Kaixin Wang, and Jiang Bian. Co-evolving latent action world models, 2025.

[39] Zizhao Wang, Chang Shi, Jiaheng Hu, Kevin Rohling, Roberto Martín-Martín, Amy Zhang, and Peter Stone. Factored latent action world models, 2026.

[40] Shihan Wu, Xu Luo, Ji Zhang, Junlin Xie, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Policy contrastive decoding for robotic foundation models, 2025.

[41] Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, Ping Luo, Xiangyu Yue, and Hongyang Li. Rise: Self-improving robot policy with compositional world model, 2026.

[42] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Ayzaan Wahid, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In *CORL*, 2020.

[43] Tony Z. Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. In *CORL*, 2024.

[44] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation, 2024.

[45] Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation. In *ICRA*, 2025.