

World4Omni: A Zero-Shot Framework from Image Generation World Model to Robotic Manipulation

Haonan Chen^{*1,2} Bangjun Wang^{*3} Jingxiang Guo^{*1} Tianrui Zhang⁴
Yiwen Hou¹ Xuchuan Huang⁵ Chenrui Tie¹ Lin Shao^{1,2}

Abstract

Improving data efficiency and generalization in robotic manipulation remains a core challenge. We propose a novel framework that leverages a pre-trained multimodal image-generation model as a world model to guide policy learning. By exploiting its rich visual-semantic representations and strong generalization across diverse scenes, the model generates open-ended future state predictions that inform downstream manipulation. Coupled with zero-shot low-level control modules, our approach enables general-purpose robotic manipulation without task-specific training. Experiments in both simulation and real-world environments demonstrate that our method achieves effective performance across a wide range of manipulation tasks with no additional data collection or fine-tuning. Supplementary materials are available on our website: <https://world4omni.github.io/>.

1. Introduction

General-purpose embodied intelligence has long been a central aspiration in AI and robotics research (Xu et al., 2024; Liu et al., 2024b; Xian et al., 2023), aiming to develop versatile robotic agents capable of handling diverse real-world tasks. Despite recent advancements, generalization remains a critical challenge. To perform manipulation tasks, robots must perceive environments, interpret complex instructions, and execute appropriate actions. However, variability in environments, tasks, objects, and robot embodiments significantly impacts robotic performance and poses considerable difficulties for generalization (Peng et al., 2018; Zhang & Yang, 2021; Wang et al., 2024). Many existing manipulation methods exhibit strong performance when operating in

scenarios similar to their training environments, yet they frequently fail in unseen contexts (Peng et al., 2018; Roy et al., 2021; Kroemer et al., 2019). Addressing these limitations typically involves two primary pathways: (1) increasing the volume and diversity of training data, and (2) developing techniques that enhance data efficiency.

Foundation models, pretrained on massive datasets, have significantly enhanced generalization in robotic tasks (Huang et al., 2023; Yang et al., 2025; Li et al., 2024). One paradigm involves training end-to-end models that directly map visual observations and language instructions to low-level actions (Brohan et al., 2022; 2023; Kim et al., 2024; Liu et al., 2024a). However, collecting robot action data remains costly and time-consuming (Yang et al., 2025). Consequently, even the largest robotics datasets (O’Neill et al., 2024) are dwarfed by Internet-scale text and image corpora (Schuhmann et al., 2022), which limits these methods’ capacity to generalize to novel tasks and scenarios (Brohan et al., 2022; 2023; Kim et al., 2024; Liu et al., 2024a). An alternative paradigm adopts a hierarchical structure, leveraging Large Language Models (LLMs) and Vision-Language Models (VLMs), pretrained on extensive textual and visual data, to perform high-level planning and prediction before interfacing with low-level action modules (Black et al., 2023; Zhen et al., 2024; Bharadhwaj et al., 2024; Huang et al., 2024). Although LLMs and VLMs improve high-level generalization, their text-based outputs restrict flexibility when integrating with low-level action models. Prior work relied on predefined low-level skill libraries, limiting generalization to unseen tasks (Ahn et al., 2022; Liang et al., 2023; Driess et al., 2023). Subsequent approaches introduced intermediate representations, yet their low-level policies still depend on additional action-labeled data, constraining overall generalizability (Black et al., 2023; Zhen et al., 2024; Wu et al., 2024).

The ability of pre-trained foundation models to generate images has recently attracted widespread attention (Yan et al., 2025; Guo & Chen, 2025; Chen et al., 2025b). We found multimodal large-scale models trained on extensive web text–image data exhibit strong generalization across diverse scenarios, suggesting their suitability as a world model for

^{*}Equal contribution ¹School of Computing, National University of Singapore ²NUS Guangzhou Research Translation and Innovation Institute ³School of computer science, Shanghai Jiao Tong University ⁴Institute for Interdisciplinary Information Science, Tsinghua University ⁵Yuanpei College, Peking University. Correspondence to: Lin Shao <linshao@nus.edu.sg>.

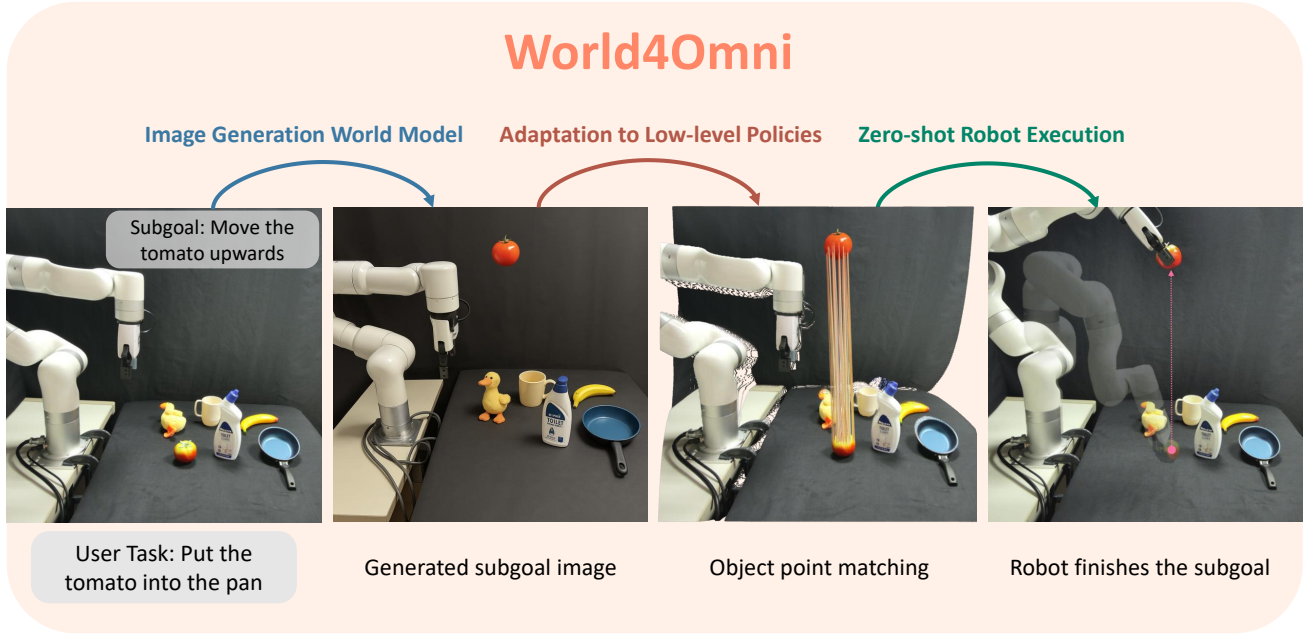


Figure 1. Overview of the World4Omni framework. We propose World4Omni, which leverages a pretrained multimodal image-generation model as a world model to guide low-level policy. Task instructions are decomposed into subtasks, each of which is fed into the world model along with the current scene image to generate a subgoal image depicting the scene after completing the current subtask. Predicted future images can be transformed into point clouds, enabling the high-level world model to adapt across different low-level policies. Object point matching validates the plausibility of predicted future images and enables their translation into concrete robot actions. Finally, a low-level policy is used to move the object from its initial position to its target position.

robotic manipulation. Prior studies have shown that world models can substantially enhance data efficiency, thereby alleviating the generalization gaps caused by data scarcity in robotics (Hafner et al., 2019; 2020; Wu et al., 2023). Other methods employ video-generation models as world models (Wu et al., 2024; Mendonca et al., 2023; Zhu et al., 2025; Gao et al., 2024; Soni et al., 2024); yet, generating future videos requires far greater temporal consistency than generating future images. Current large pre-trained video-generation models fail to achieve zero-shot generalization in robotic manipulation tasks. As a result, most of these approaches still require additional task-specific training.

In this work, we employ a pre-trained foundation model as a world model to generate images depicting future object states. To mitigate inconsistencies in image outputs, we introduce a VLM as a Reflection Agent, which evaluates and refines these generated images. Additionally, we propose a Task Planner Agent that decomposes tasks into sequential subtasks, enhancing reasoning ability for long-horizon tasks. Thanks to the rich representational capacity of images, the future scenes generated by the world model can be converted into other modalities. For example, they can subsequently be transformed into point clouds via single-view depth estimation techniques (Yang et al., 2024a). As a result, our framework supports diverse input modalities for

low-level modules—including current and predicted RGB images, point clouds, and structured representations derived from them (e.g., keypoints or object transformations). We evaluate our approach on representative manipulation tasks, comparing its zero-shot generalization performance with other hierarchical methods and across different paradigms.

Overall, our main contributions are as follows:

- We introduce a novel framework, **World4Omni**, capable of zero-shot, cross-embodiment generalization across diverse robotic manipulation tasks without any additional training.
- We employ a pre-trained large-scale multimodal image-generation model as a world model, incorporating an agent-based collaborative reflection process to iteratively refine imagined future scenes, thereby generating more plausible and consistent subgoal images.
- Our framework supports plug-and-play integration of low-level modules designed for different input modalities, showcasing its versatility and strong adaptability.
- We demonstrate the zero-shot generalization and cross-embodiment capabilities of our framework on diverse simulated and real-world robotic manipulation tasks, achieving favorable results across the evaluations.

2. Related Work

2.1. World Models for Robotic Manipulation

Early work on world models for robotic manipulation primarily focused on learning visual dynamics directly from raw pixel observations to predict future frames (Finn & Levine, 2017; Ebert et al., 2018). Subsequently, latent-space world models were introduced to encode the underlying physical dynamics compactly. For instance, Dreamer and its variants (Hafner et al., 2019; 2020; 2023; Wu et al., 2023) learn internal latent-state representations and optimize robot behavior by simulating or “imagining” future trajectories. These latent-space models enhance data efficiency by augmenting limited real-world data with imagined experiences. Recent studies have also explored video-generation models as world models (Mendonca et al., 2023; Wu et al., 2024; Zhu et al., 2025; Gao et al., 2024). Although promising, such models require high temporal consistency, and existing pre-trained, large-scale video-generation models often fail to generalize effectively to novel scenarios (Wu et al., 2024; Mendonca et al., 2023; Zhu et al., 2025; Gao et al., 2024; Soni et al., 2024), thus limiting their practical application in diverse robotic manipulation tasks (Rigter et al., 2024; Soni et al., 2024). In this work, we use pre-trained, large-scale image-generation models as world models. By predicting only essential subgoal images at key frames, our method avoids the temporal consistency issues faced by video-generation approaches and achieves robust generalization across various robotic manipulation scenarios.

2.2. Reflection in Foundation Models

Reflection mechanisms, which enable generative models to iteratively critique and refine their outputs, have recently attracted growing attention as a promising approach for enhancing robotic manipulation capabilities. In generative modeling, recent studies demonstrate that incorporating self-feedback or iterative critiques substantially improves the quality and coherence of generated outputs (Shinn et al., 2023; Madaan et al., 2023; Raman et al., 2024). Notable examples include CritiqueLLM (Ke et al., 2023) and Idea2Img (Yang et al., 2024c), which showcase how reflective feedback loops facilitate progressive refinement and correction of initial predictions. Extending these reflective approaches into robotics, several recent frameworks (Feng et al., 2025; Yang et al., 2024b) integrate self-reflection into robotic task planning and action execution, enabling robotic agents to dynamically identify and correct errors, thereby progressively enhancing their performance during tasks. Moreover, additional studies have advanced this concept by incorporating multimodal reflection mechanisms, effectively bridging high-level cognitive reasoning with low-level motor control adjustments. This multimodal integration significantly improves robot robustness and adaptability,

enabling robots to better manage uncertainties and effectively generalize across diverse manipulation scenarios and real-world conditions (Liu et al., 2025; Xiong et al., 2024; Xia et al., 2025).

2.3. Foundation Models Paradigms for Robotic Manipulation

Recent advances in foundation models have significantly influenced robotics, especially in robotic manipulation tasks, by leveraging LLMs and VLMs for high-level planning and decision-making (Firoozi et al., 2023; Li et al., 2025; Chen et al., 2025a; Tie et al., 2025; Patel et al., 2025). Current research can be broadly categorized into three main paradigms. The first paradigm integrates foundation models into end-to-end frameworks that map visual and linguistic inputs directly to continuous low-level robot actions. Notable methods include RT-1 (Brohan et al., 2022), RT-2 (Brohan et al., 2023), OpenVLA (Kim et al., 2024), and RDT-1B (Liu et al., 2024a). These methods are trained jointly on robotic action data and vision-language corpora to predict robot motions directly from image and language inputs. However, the high cost and time required to collect robot action data severely constrain dataset size, making robust zero-shot generalization under low-data conditions challenging. The second paradigm employs foundation models to guide robotic execution by linking high-level instructions to predefined low-level skills. Approaches such as SayCan (Ahn et al., 2022), PALM-E (Driess et al., 2023), and Code as Policies (Liang et al., 2023) utilize LLM outputs combined with skill libraries or executable code generation to bridge high-level planning and robotic actions. However, these methods often struggle to generalize predefined skill sets. The third paradigm introduces intermediate visual representations or subgoals to enhance generalization and task execution. Methods such as ReKep (Huang et al., 2024), SuSIE (Black et al., 2023), 3D-VLA (Zhen et al., 2024), and Gen2Act (Bharadhwaj et al., 2024) use foundation models to generate intermediate goals, like keypoints, subgoal images, or demonstration videos, that guide robotic policies, thus improving adaptability to novel scenarios.

3. Method

Our framework is illustrated in Figure 2. In this section, we provide a detailed explanation of problem formulation (Sec. 3.1), agent collaboration (Sec. 3.2), reflective world model (Sec. 3.3), and low-level policy (Sec. 3.4).

3.1. Problem Formulation

Given a single-view RGB image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, a point cloud $\mathcal{P} \in \mathbb{R}^{N \times 3}$, and a natural language task description \mathcal{L} , the objective is to generate an action sequence $\{\mathbf{a}\}_i$ that completes the manipulation task described in \mathcal{L} .

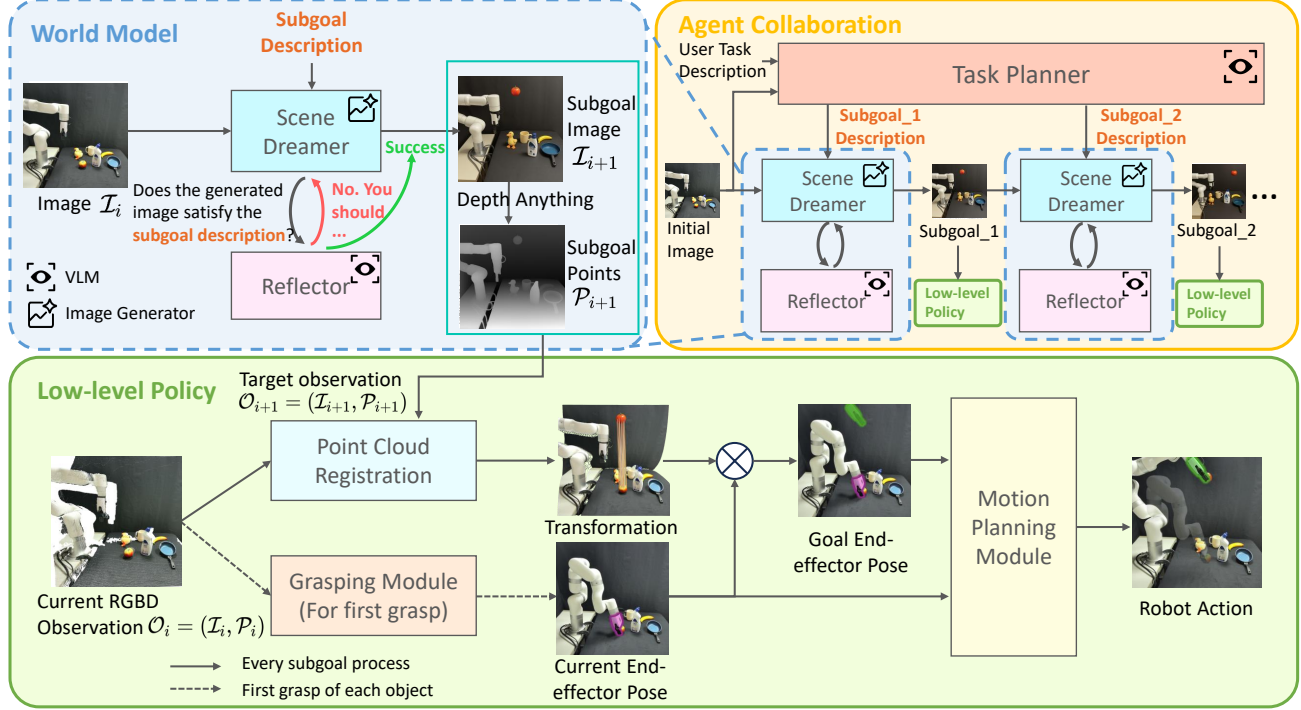


Figure 2. **An instantiation of our framework.** Agent Collaboration involves the Task Planner, Scene Dreamer, and Reflector agents working together (yellow). The Scene Dreamer and Reflector together form the Reflective World Model, which produces subgoal images and point clouds (blue). The Low-level Policy consumes the current and goal observations and outputs the robot actions (green).

The **Task Planner Agent** takes an RGB image \mathcal{I} and a task description \mathcal{L} as input and outputs a sequence of subtask descriptions $\{\mathcal{L}_i\}$.

The **World Model** receives an RGB image \mathcal{I}_i and a subtask description \mathcal{L}_i to produce a subgoal image $\mathcal{I}_{i+1} \in \mathbb{R}^{H \times W \times 3}$ and a subgoal depth map $\mathcal{D}_{i+1} \in \mathbb{R}^{H \times W \times 1}$. The resulting future point cloud \mathcal{P}_{i+1} is obtained by back-projecting \mathcal{D}_{i+1} .

The **Low-level Policy** is provided with the current observation $\mathcal{O}_i = (\mathcal{I}_i, \mathcal{P}_i)$ and the future observation $\mathcal{O}_{i+1} = (\mathcal{I}_{i+1}, \mathcal{P}_{i+1})$, and outputs a sequence of actions $\{\mathbf{a}\}_i$ to drive the system from \mathcal{O}_i to \mathcal{O}_{i+1} .

Notably, the low-level policy may use any non-empty subset of the modalities from the current observation $\mathcal{O}_i = (\mathcal{I}_i, \mathcal{P}_i)$ (i.e., \mathcal{I}_i , \mathcal{P}_i , or both) and any non-empty subset from the target observation $\mathcal{O}_{i+1} = (\mathcal{I}_{i+1}, \mathcal{P}_{i+1})$ (i.e., \mathcal{I}_{i+1} , \mathcal{P}_{i+1} , or both) to produce the action sequence $\{\mathbf{a}\}_i$. This shows our framework is compatible with a variety of different low-level policy settings.

3.2. Agent Collaboration

Given the user’s task description \mathcal{L} and a scene image \mathcal{I} , we employ a VLM (GPT-o4-mini-high) as the Task Planner

Agent to decompose the task into a sequence of subtask descriptions $\{\mathcal{L}_i\}$ in text form. For example, given the initial scene image and the instruction ”Put the tomato in the pan”, the Task Planner Agent might decompose the task into the following subtasks: (1) \mathcal{L}_0 : ”Move the tomato vertically upward”; (2) \mathcal{L}_1 : ”Move the tomato horizontally to the right, positioning it above the pan”; (3) \mathcal{L}_2 : ”Move the tomato downward into the pan”. These subtasks are then executed sequentially until the overall task is completed.

At the start, we input the initial image \mathcal{I}_0 (specifically, $\mathcal{I}_0 = \mathcal{I}$) along with its corresponding subtask description \mathcal{L}_0 into the Reflective World Model. The Reflective World Model consists of a Scene Dreamer Agent and a Reflector Agent (detailed in Sec. 3.3), and produces a predicted future scene image \mathcal{I}_1 . We refer to each such predicted future scene image as a subgoal image. By the same process, for each input image \mathcal{I}_i together with its subtask description \mathcal{L}_i , the Reflective World Model outputs the subgoal image \mathcal{I}_{i+1} . As illustrated in Figure 2, the initial desktop image and the first subtask instruction \mathcal{L}_0 are input to the World Model, which outputs a subgoal image showing the tomato moving upward. This subgoal image, together with the next subtask instruction \mathcal{L}_1 , is then fed into the World Model to produce the next subgoal image of the tomato moving to the right, and this process continues until the task is complete.

Subgoal images serve two roles: (1) they can be used directly by a low-level policy or converted into another modality for a low-level policy (see Sec. 3.3); and (2) they provide the input for the next Scene Dreamer Agent to generate the subsequent subgoal image. By using subgoal images rather than ground-truth scene images as input for subsequent processing, we avoid issues where objects of interest are occluded by the robot arm or gripper, allowing the Reflective World Model to output more consistent images.

3.3. Reflective World Model

The Reflective World Model consists of a Scene Dreamer Agent, which leverages a large-scale pre-trained image-generation model, and a Reflector Agent, which is built on a VLM. The Scene Dreamer Agent receives the scene image \mathcal{I}_i and the subtask description \mathcal{L}_i from the Task Planner, then employs GPT-4o to generate an image of the future scene \mathcal{I}_{i+1} . The Reflector Agent employs GPT-o4-mini-high to understand both the generated image and the subtask semantics, evaluating whether the output of the Scene Dreamer Agent is consistent. If the image passes this check, the Reflector emits a success signal; if not, it produces a revised prompt to steer the Scene Dreamer toward a more accurate generation.

Taking the future scene generation in Fig. 2 as an example, if the output of Scene Dreamer Agent shows incorrect movement of the target object or disregards the surrounding context, the Reflector Agent issues a revised prompt \mathcal{L}'_i and submits it along with the current scene image \mathcal{I}_i to the Scene Dreamer Agent. Then the Scene Dreamer Agent generates a new image \mathcal{I}_{i+1} . This reflective loop mitigates hallucinations and goal inconsistencies in image outputs of the Scene Dreamer Agent. Finally, the resulting scene images \mathcal{I}_{i+1} can be converted into depth maps \mathcal{P}_{i+1} using Depth-Anything (Yang et al., 2024a), allowing flexible support for different inputs from low-level models.

In robotic manipulation tasks, we concentrate on the object of interest; accordingly, our generated images are object-centric. Although image-generation models may introduce inconsistencies in background elements, we disregard these artifacts and concentrate solely on the target object. To ensure a clean, focused representation for the low-level policy, we use Grounded SAM (Ren et al., 2024) to segment out the object of interest.

3.4. Low-level Policy

In our framework, the high-level model provides the low-level policy with both the current observation $\mathcal{O}_i = (\mathcal{I}_i, \mathcal{P}_i)$ and the target observation $\mathcal{O}_{i+1} = (\mathcal{I}_{i+1}, \mathcal{P}_{i+1})$, forming the complete set of available inputs. The low-level policy may then select any non-empty subset of these observations as its input. Our framework is adaptive to different low-level

policies—any policy that can operate on these inputs can be seamlessly integrated. Specifically, as an instantiation of the low-level policy of our framework, we employ the Grasping+Planning approach, which comprises three core components: Point Cloud Registration, a Grasping Module, and a Motion Planning Module.

Point Cloud Registration. We adopt GeoAware-SC (Zhang et al., 2024) as our principal module for point-cloud registration, exploiting its geometry-aware semantic correspondence module to establish dense, per-pixel matches between the initial and subgoal images. Concurrently, we apply Depth-Anything (Yang et al., 2024a) to both views to infer high-resolution depth maps. To delineate object extents within both scenes, we integrate Grounded SAM (Ren et al., 2024), generating robust segmentation masks for all salient entities. Finally, given the fused semantic correspondences and their associated metric depths, we lift the depth into 3D point clouds and employ the Umeyama algorithm (Umeyama, 1988) to estimate the optimal rigid transformation that aligns the initial and subgoal poses.

Grasping Module. Our low-level policy framework employs pre-trained GraspNet (Fang et al., 2020) as its grasping module, which takes a point cloud input and generates top-K grasp poses in the camera frame in an end-to-end manner. For implementation, we first use Grounded SAM (Ren et al., 2024) to generate masks and segment target object points from the original point cloud. We then filter the GraspNet-generated poses, keeping only those within a distance threshold of these target points, and select the one with the highest score.

Motion Planning Module. The Motion Planning Module generates a sequence of control commands that drive the manipulator from its initial pose to the target pose. In the simulated environment, we leverage a sample-based motion planning module built in the simulator, which can interpolate trajectories from the initial to the target pose. In our real-world experiments, we employ optimization-based motion planning, optimizing a cost function over candidate trajectories and selecting the one with the lowest cost.

4. Experiment

4.1. Experimental Setup

Simulation. Simulation experiments were carried out in RLBench (James et al., 2020) using a Franka Emika Panda 7-DoF arm and several RGB-D cameras. The robot arm is fixed to the tabletop, and the objects are randomly placed on the table. At the start of every trial, no objects are held.

Some typical robot manipulation tasks in RLBench are selected. In Table 1 and Table 2, Tasks 1–4 correspond to OpenWineBottle, TakePlateOffColoredDishRack, Take-

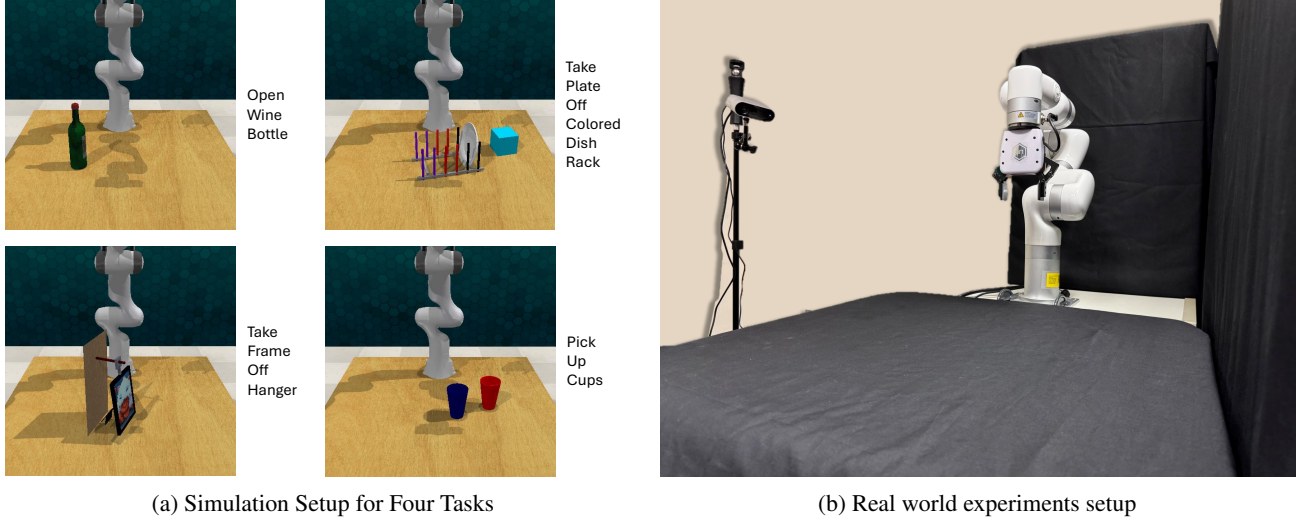


Figure 3. Experiment setup in simulation and real world.

FrameOffHanger, and PickUpCups, respectively (Figure 3a). To ensure consistency with the baselines, we assign each RL Bench task a random seed from 1 to 5, corresponding to five different arrangements for the task. For each seed, we conduct 10 trials, resulting in a total of 50 runs for each task to robustly assess performance.

Real World. Our real-world experimental setup, as depicted in Figure 3b, comprises a 7-DoF UFACTORY X-ARM 7 and an Orbbec Femto Bolt RGB-D camera. At the start of each trial, the robot does not grasp any object. We evaluate our pipeline on two real-world robotic manipulation tasks. 1) Move the tomato into the pan. 2) Take the plate off the rack, thereby demonstrating its practical applicability.

4.2. Baselines

Representative methods from different paradigms are selected as our baselines.

- For cross-paradigm comparison, we adopt the end-to-end OpenVLA (Kim et al., 2024) approach as a baseline.
- For intra-paradigm comparisons, our baseline is SuSIE (Black et al., 2023), which leverages a pre-trained image-editing model.

We also compare various high-level world models and low-level policies within our framework.

- For world models, we compare GPT-4o (Hurst et al., 2024), DALL-E 3 (Betker et al., 2023), and Gemini 2.5 Pro (Team et al., 2023) for their image-generation performance as well as Sora (Liu et al., 2024c; Brooks et al., 2024) for video generation.

- For zero-shot low-level policies, we compare our Grasping+Planning method with the Octo (Team et al., 2024) foundation model that conditions on both initial and goal images.

Neither the baselines nor our method undergoes any additional training in RL Bench or on real-world setups; all comparisons are conducted under zero-shot settings.

4.3. Cross-Paradigm Comparison

Table 1 presents the results of our cross-paradigm and intra-paradigm evaluations. As described in Sec. 4.1, each method was tested on 50 trials for the RL Bench tasks OpenWineBottle, TakePlateOffColoredDishRack, and PickUpCups. In the table, results are reported as "number of successes/number of trials", and the rightmost column shows the average success rate across tasks.

For the end-to-end OpenVLA approach, the average success rate was 0%, indicating its inability to execute RL Bench tasks in a zero-shot setting. This underscores its limited ability to generalize, making it challenging to transfer to unseen scenes and tasks. Moreover, its fully integrated, closed-box design offers no observable internal states, preventing us from pinpointing the exact causes of failure.

4.4. Intra-Paradigm Comparison

For the hierarchical model SuSIE, the average success rate on RL Bench is 0%, demonstrating that it struggle to complete these tasks in a zero-shot setting. For SuSIE, we observed severe hallucinations in its predicted images (see Fig. 4), making it impossible to generate correct goal images. For example, in the TakePlateOffColoredDishRack task, the generated image is completely different from the

Table 1. Cross-paradigm and intra-paradigm experiment results in simulation

Method	Task1	Task2	Task3	Task4	Average Success Rate
OpenVLA	0 / 50	0 / 50	0 / 50	0 / 50	0%
SUSIE	0 / 50	0 / 50	0 / 50	0 / 50	0%
Ours	10 / 50	20 / 50	10 / 50	30 / 50	35%

Table 2. Zero-shot low-level policy evaluation results in simulation

Method	Task1	Task2	Task3	Task4	Average Success Rate
Octo	0 / 50	0 / 50	0 / 50	0 / 50	0%
Grasping+Planning	10 / 50	20 / 50	10 / 50	30 / 50	35%

original image. Consequently, its low-level policy lacked reliable targets and failed to complete the task.

By contrast, the image generation world model in our framework produces more accurate future images that can effectively guide the zero-shot low-level policy (see Fig. 4). The results in Table 1 demonstrate that our framework achieves strong generalization, completing manipulation tasks without any additional training.

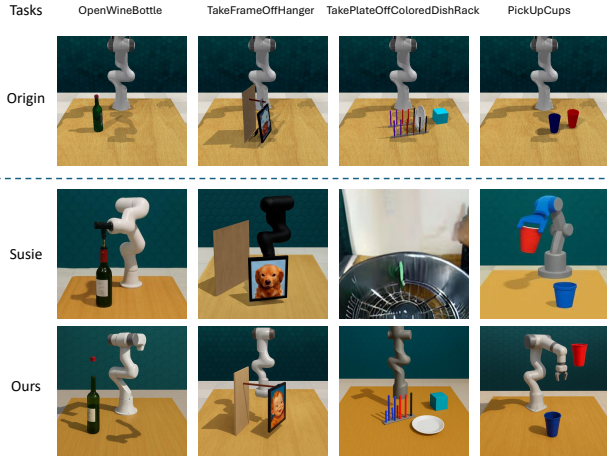


Figure 4. Intra-Paradigm Comparison of generated future images.

4.5. Image-Generation World Model Comparison

We compared several widely used large multimodal image-generation models. As illustrated in Figure 5, the leftmost panel shows the initial input image. We provided each model with the same subgoal description and conducted a qualitative comparison of their generated images. The top row illustrates the generated simulation images for the subgoal "Take Plate Off Colored Dish Rack." The bottom row illustrates the generated real-world images for the sub-

goal "Move the tomato upwards." We focused exclusively on the correctness of the target object's placement, without considering the consistency of other scene elements.

We observed that GPT-4o's outputs often exhibit stylistic variation while still positioning the object at the intended location. Gemini preserves the original style more faithfully, yet tends to render multiple copies of the target object. The video-generation model Sora suffers from pronounced hallucinations, resulting in drastic scene alterations and poor temporal coherence. Lastly, DALL-E 3 demonstrates limited understanding of scene structure and spatial relationships, resulting in incorrect object placements in the environment.

4.6. Zero-shot Low-level Policy Comparison

To compare low-level policies, we evaluated Octo, which takes both the current and predicted future images as input and outputs the actions needed to move the robot toward the goal view. Our experiments show that Octo is unable to zero-shot generalize to the Franka manipulator in RL Bench, resulting in task failures.

In contrast, our Grasping+Planning approach achieved a success rate of 35%, surpassing Octo. This demonstrates that the Grasping+Planning module is capable of zero-shot generalization. Most failures in this approach result from the GraspNet module's inability to generate a successful grasp, leading to task failure.

4.7. Real World Experiments

Our real-world experiment is illustrated in Figure 6. The left panel illustrates how our World4Omni framework uses GPT-4o as the world model to generate future scene images. The goal of the task is to move the tomato into the pan. As depicted, the world model successfully produces the goal image of the task. The right panel of Figure 6 shows the task execution, where the robot completes the task in a zero-shot

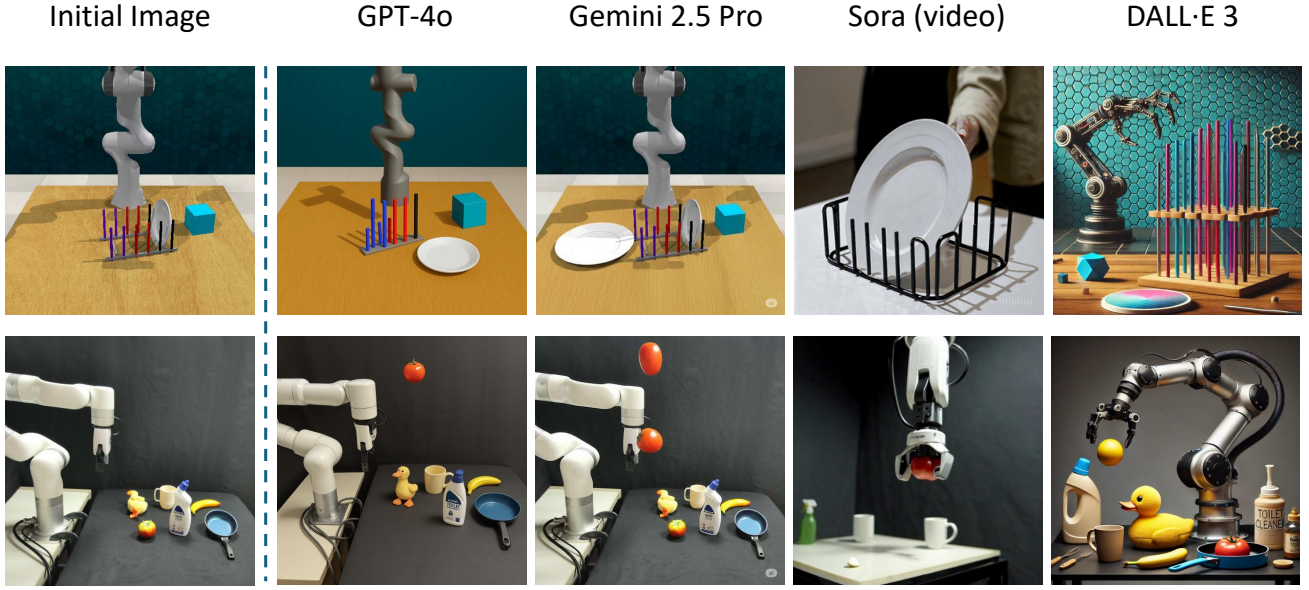


Figure 5. Image generation world model comparison.

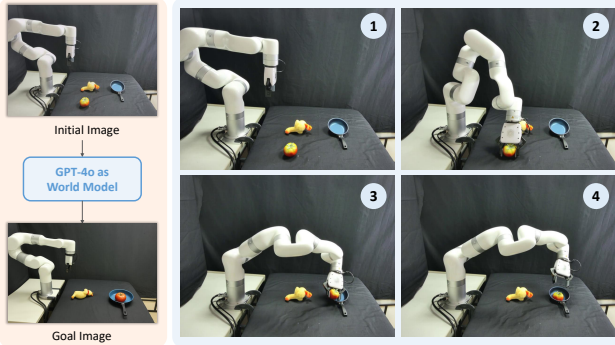


Figure 6. Real World Experiment.

manner without any additional training. This showcases the zero-shot generalization capability of our framework.

5. Limitation and Future Work

Although the World4Omni framework can execute a variety of robotic manipulation tasks in a zero-shot, cross-embodiment fashion, it makes several trade-offs to achieve this level of generalization. Using the large pre-trained image generation model improves generalization, but it sometimes fails to maintain spatial accuracy, making precise operations such as insertion challenging. Inconsistencies in the generated images further complicate execution, and occlusions from the gripper or robot arm impede point-cloud matching, making closed-loop control difficult. Moreover, the foundation grasping module struggles to perform func-

tional grasps on articulated and deformable objects. Although high-level models can generate reasonable future images, the limitations of the low-level policy prevent these tasks from being completed successfully. These findings highlight the need for future research to develop more powerful and general-purpose low-level foundation models.

6. Conclusion

In this work, we introduce **World4Omni**, a hierarchical robot manipulation framework that uses images as intermediate representations. The framework uses a Reflective World Model to generate future scene images and point clouds, where a VLM provides reflective feedback to refine the image quality produced by the pre-trained image-generation model. A zero-shot low-level policy then consumes current and predicted future images (or their corresponding point clouds) to produce robot actions without any additional training. Our cross-paradigm and intra-paradigm evaluations show that World4Omni surpasses representative methods in zero-shot generalization. Moreover, we achieve success in both simulation and real-world settings without any additional training, demonstrating strong generalization and cross-embodiment capabilities. We demonstrate that, by using images generated by foundation models as intermediate representations and executing low-level policies with no additional training, robots can achieve both strong generalization and cross-embodiment across diverse manipulation tasks. This result points the field toward a promising path for realizing general-purpose embodied intelligence.

Impact Statement

This paper presents work whose goal is to advance the field of Embodied AI. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by the School of Computing, National University of Singapore, and by the NUS Guangzhou Research Translation and Innovation Institute.

References

- Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3): 8, 2023.
- Bharadhwaj, H., Dwibedi, D., Gupta, A., Tulsiani, S., Dorsch, C., Xiao, T., Shah, D., Xia, F., Sadigh, D., and Kirmani, S. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.
- Black, K., Nakamoto, M., Atreya, P., Walke, H., Finn, C., Kumar, A., and Levine, S. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.
- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. *OpenAI Blog*, 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>. Online; accessed 16 May 2025.
- Chen, H., Li, J., Wu, R., Liu, Y., Hou, Y., Xu, Z., Guo, J., Gao, C., Wei, Z., Xu, S., et al. Metafold: Language-guided multi-category garment folding framework via trajectory generation and foundation model. *arXiv preprint arXiv:2503.08372*, 2025a.
- Chen, S., Bai, J., Zhao, Z., Ye, T., Shi, Q., Zhou, D., Chai, W., Lin, X., Wu, J., Tang, C., et al. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*, 2025b.
- Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488, Honolulu, Hawaii, USA, 2023. PMLR. URL <https://proceedings.mlr.press/v202/driess23a.html>.
- Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., and Levine, S. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.
- Fang, H.-S., Wang, C., Gou, M., and Lu, C. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11444–11453, 2020.
- Feng, Y., Han, J., Yang, Z., Yue, X., Levine, S., and Luo, J. Reflective planning: Vision-language models for multi-stage long-horizon robotic manipulation. *arXiv preprint arXiv:2502.16707*, 2025.
- Finn, C. and Levine, S. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2786–2793. IEEE, 2017.
- Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K., et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, pp. 02783649241281508, 2023.
- Gao, C., Zhang, H., Xu, Z., Cai, Z., and Shao, L. Flip: Flow-centric generative planning for general-purpose manipulation tasks. *arXiv preprint arXiv:2412.08261*, 2024.
- Guo, J. and Chen, H. Can gpt-4o image generation unlock new potential in robotic manipulation? *TechRxiv*, April 2025. doi: 10.36227/techrxiv.174535631.14854732/

- v1. URL <http://dx.doi.org/10.36227/techrxiv.174535631.14854732/v1>. Preprint.
- Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.
- Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.
- Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.
- Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., and Fei-Fei, L. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.
- Huang, W., Wang, C., Li, Y., Zhang, R., and Fei-Fei, L. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.
- Ke, P., Wen, B., Feng, Z., Liu, X., Lei, X., Cheng, J., Wang, S., Zeng, A., Dong, Y., Wang, H., et al. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. *arXiv preprint arXiv:2311.18702*, 2023.
- Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sankeketi, P., et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- Kroemer, O., Niekum, S., and Konidaris, G. A review of robot learning for manipulation: Challenges. *Representations, and Algorithms*, pp. 82, 2019.
- Li, D., Jin, Y., Sun, Y., Yu, H., Shi, J., Hao, X., Hao, P., Liu, H., Sun, F., Zhang, J., et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024.
- Li, L., Yuan, L., Liu, P., Jiang, T., and Yu, Y. Llm-assisted semantically diverse teammate generation for efficient multi-agent coordination. In *Proceedings of the Forty-second International Conference on Machine Learning*, 2025.
- Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.
- Liu, J., Li, C., Wang, G., Li, X., Chen, S., Xiong, C., Ge, J., Zhou, K., and Zhang, S. Self-corrected multimodal large language model for robot manipulation and reflection. In *International Conference on Learning Representations (ICLR)*, September 2025. URL <https://openreview.net/forum?id=TLWbNfbkxj>. Withdrawn Submission.
- Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., and Zhu, J. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024a.
- Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., and Lin, L. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024b.
- Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024c.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- Mendonca, R., Bahl, S., and Pathak, D. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023.
- O’Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandekar, A., Jain, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.
- Patel, S., Yin, X., Huang, W., Garg, S., Nayyeri, H., Fei-Fei, L., Lazebnik, S., and Li, Y. A real-to-sim-to-real approach to robotic manipulation with vlm-generated iterative keypoint rewards. *arXiv preprint arXiv:2502.08643*, 2025.
- Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on*

- robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.
- Raman, S. S., Cohen, V., Idrees, I., Rosen, E., Mooney, R., Tellex, S., and Paulius, D. Cape: Corrective actions from precondition errors using large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14070–14077. IEEE, 2024.
- Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.
- Rigter, M., Gupta, T., Hilmkil, A., and Ma, C. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.
- Roy, N., Posner, I., Barfoot, T., Beaudoin, P., Bengio, Y., Bohg, J., Brock, O., Depatie, I., Fox, D., Koditschek, D., et al. From machine learning to robotics: Challenges and opportunities for embodied intelligence. *arXiv preprint arXiv:2110.15245*, 2021.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- Soni, A., Venkataraman, S., Chandra, A., Fischmeister, S., Liang, P., Dai, B., and Yang, S. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024.
- Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.
- Tie, C., Sun, S., Zhu, J., Liu, Y., Guo, J., Hu, Y., Chen, H., Chen, J., Wu, R., and Shao, L. Manual2skill: Learning to read manuals and acquire robotic skills for furniture assembly using vision-language models. *arXiv preprint arXiv:2502.10090*, 2025.
- Umeyama, S. An eigendecomposition approach to weighted graph matching problems. *IEEE transactions on pattern analysis and machine intelligence*, 10(5):695–703, 1988.
- Wang, T., Bhatt, D., Wang, X., and Atanasov, N. Cross-embodiment robot manipulation skill transfer using latent space alignment. *arXiv preprint arXiv:2406.01968*, 2024.
- Wu, J., Yin, S., Feng, N., He, X., Li, D., Hao, J., and Long, M. ivideopt: Interactive videopts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.
- Wu, P., Escontrela, A., Hafner, D., Abbeel, P., and Goldberg, K. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023.
- Xia, W., Feng, R., Wang, D., and Hu, D. Phoenix: A motion-based self-reflection framework for fine-grained robotic action correction. *arXiv preprint arXiv:2504.14588*, 2025.
- Xian, Z., Gervet, T., Xu, Z., Qiao, Y.-L., Wang, T.-H., and Wang, Y. Towards generalist robots: A promising paradigm via generative simulation. *arXiv preprint arXiv:2305.10455*, 2023.
- Xiong, C., Shen, C., Li, X., Zhou, K., Liu, J., Wang, R., and Dong, H. Aic mllm: Autonomous interactive correction mllm for robust robotic manipulation. *arXiv preprint arXiv:2406.11548*, 2024.
- Xu, Z., Wu, K., Wen, J., Li, J., Liu, N., Che, Z., and Tang, J. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*, 2024.
- Yan, Z., Ye, J., Li, W., Huang, Z., Yuan, S., He, X., Lin, K., He, J., He, C., and Yuan, L. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025.
- Yang, J., Tan, W., Jin, C., Yao, K., Liu, B., Fu, J., Song, R., Wu, G., and Wang, L. Transferring foundation models for generalizable robotic manipulation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1999–2010. IEEE, 2025.
- Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.
- Yang, Z., Garrett, C., Fox, D., Lozano-Pérez, T., and Kaelbling, L. P. Guiding long-horizon task and motion planning with vision language models. *arXiv preprint arXiv:2410.02193*, 2024b.

- Yang, Z., Wang, J., Li, L., Lin, K., Lin, C.-C., Liu, Z., and Wang, L. Idea2img: Iterative self-refinement with gpt-4v for automatic image design and generation. In *European Conference on Computer Vision*, pp. 167–184. Springer, 2024c.
- Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., and Yang, M.-H. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Zhang, Y. and Yang, Q. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.
- Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., and Gan, C. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.
- Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., and Gupta, A. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.