Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Act2Goal: From World Model To General Goal-conditioned Policy

Pengfei Zhou[1,*], Liliang Chen[1,*], Shengcong Chen[1], Di Chen[1],
Wenzhi Zhao[1], Rongjun Jin[1], Guanghui Ren[1], Jianlan Luo[1,†]
[1]Agibot Research
[*]Equal contribution [†]Corresponding author

***Abstract*—Specifying robotic manipulation tasks in a manner that is both expressive and precise remains a central challenge. While visual goals provide a compact and unambiguous task specification, existing goal-conditioned policies often struggle with long-horizon manipulation due to their reliance on single-step action prediction without explicit modeling of task progress. We propose Act2Goal, a general goal-conditioned manipulation policy that integrates a goal-conditioned visual world model with multi-scale temporal control. Given a current observation and a target visual goal, the world model generates a plausible sequence of intermediate visual states that captures long-horizon structure. To translate this visual plan into robust execution, we introduce Multi-Scale Temporal Hashing, which decomposes the imagined trajectory into dense proximal frames for fine-grained closed-loop control and sparse distal frames that anchor global task consistency. The policy couples these representations with motor control through end-to-end cross-attention, enabling coherent long-horizon behavior while remaining reactive to local disturbances. Extensive experiments in both simulation and real-world environments demonstrate that Act2Goal, initialized from large-scale imitation learning, achieves strong performance across a wide range of long-horizon manipulation tasks. Beyond offline generalization, Act2Goal also supports reward-free online autonomous improvement via hindsight goal relabeling and LoRA-based finetuning, enabling rapid adaptation during deployment without external supervision. In real world experiments, it improves success rates from 30% to 90% on challenging out-of-distribution tasks within minutes of autonomous interaction, validating the effectiveness of online training as a powerful complement to offline learning. Project page: https://act2goal.github.io/**

Fig. 1: **Method Overview.** The model receives a visual goal (left), imagines how to achieve it via a goal-conditioned world model (top), and executes the planned actions in the real world (right).

I. Introduction

Learning robotic manipulation policies requires task specifications that are both expressive and precise. While natural language provides a flexible interface for high-level task descriptions, it often falls short in encoding the fine-grained spatial details needed for precise manipulation — especially in complex, multi-step scenarios. In contrast, visual goals offer a more grounded and unambiguous alternative. By directly encoding object configurations, spatial relations, and terminal constraints, they bypass the challenges of linguistic ambiguity.

Goal-conditioned policies (GCPs) map the current observation and a target visual goal directly to actions [1, 2, 3, 4]. While these methods perform well in short-horizon settings and exhibit reasonable generalization [5, 6], their performance deteriorates in long-horizon tasks. This limitation arises because most previous GCPs rely on direct action prediction without explicitly modeling task progress, intermediate feasibility, or long-term consistency [7]. This issue is further exacerbated when GCPs are trained on narrowly scoped demonstration data. Without a model of how the visual scene evolves toward the goal, such policies tend to overfit to local state–action mappings and depend heavily on dense supervision. This limitation becomes particularly pronounced in long-horizon or out-of-distribution settings, where maintaining coherent progress toward the goal requires reasoning beyond locally observed transitions.

To overcome this, GCPs must therefore explicitly model the visual dynamics required to reach a desired goal. Without such mechanism, the policy may focus narrowly on imitating local action patterns, rather than learning goal-directed behaviors that reflect progress toward the final outcome.

World models offer a promising solution to address these limitations [8, 9, 10]. Recent advances show that generative models conditioned on language instructions can predict future visual states to support planning and decision-making [11, 12,

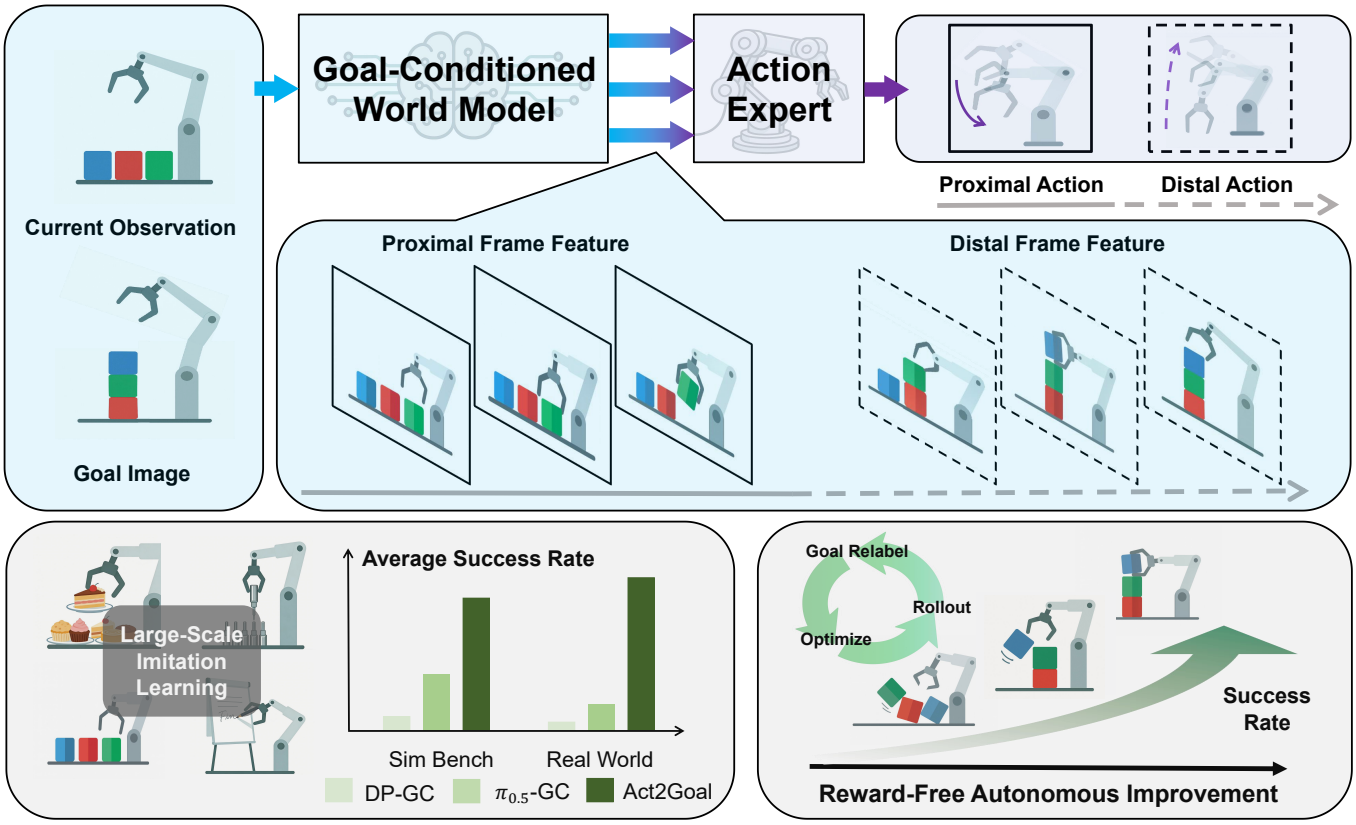

Fig. 2: **System Overview.** We propose Act2Goal, a goal-conditioned policy that integrates a visual world model with multi-scale temporal control to address long-horizon manipulation. After large-scale offline imitation learning, the model shows high performance on seen settings and strong generalization to unseen scenarios. The reward-free online autonomous improvement stage further improve model's performance through rollout-goal relabel-optimize loop.

13]. Building on this insight, we consider a goal-conditioned analogue: Goal-Conditioned World Model. Instead of predicting the future from high-level language, a goal-conditioned world model produces a plausible sequence of intermediate states bridging the gap between the current observation and a desired visual goal. This capability remedies a fundamental limitation of conventional GCPs by introducing an explicit representation of scene evolution over time, enabling a visually grounded and task-consistent trajectory toward the specified goal.

While generating a plausible visual trajectory is crucial, it alone does not ensure reliable execution. Successfully performing long-horizon manipulation tasks requires addressing a more fundamental control challenge. A general-purpose GCP must balance global consistency—maintaining fidelity to the long-term goal—with local reactivity—robustly handling perturbations and correcting errors during closed-loop execution. Full-trajectory planning provides global coherence but is brittle under deviations; short-horizon control improves robustness but often drifts off-course in extended tasks. This inherent tension poses a fundamental barrier to deploying GCPs as general-purpose robotic controllers.

To reconcile this tension, we require not only structured visual prediction, but also temporal abstraction that enables the policy to flexibly reason over both short-term feedback and long-term objectives. Thus, we propose Act2Goal, a general goal-conditioned policy that integrates a goal-conditioned world model with a novel temporal decomposition mechanism termed Multi-Scale Temporal Hashing (MSTH). MSTH decomposes the predicted visual trajectory into dense proximal frames for fine-grained control and sparse, horizon-adaptive distal frames that anchor the global plan. This multi-scale structure allows the policy to react to local disturbances while preserving long-term intent. Visual features from the world model are fused into the action expert via layer-wise cross-attention, implemented in a fully differentiable fashion, enabling intermediate scene representations to guide low-level motor control.

The proposed Act2Goal architecture not only enables robust performance through large scale offline imitation learning, but also lays the foundation for autonomous online adaptation. To this end, we incorporate a reward-free learning mechanism based on Hindsight Experience Replay (HER) [14], which relabels the model's own rollouts as goal-achieving trajectories. Combined with lightweight LoRA-based parameter updates [15], this mechanism allows the model to improve its

performance continuously and efficiently in novel scenarios, without any task-specific reward.

We evaluate the effectiveness of Act2Goal in both simulation and real-world experiments. Act2Goal achieves strong zero-shot generalization across unseen objects, spatial arrangements, and environments in a variety of manipulation tasks. It consistently outperforms competitive baselines in both in-domain and out-of-domain settings, and demonstrates rapid performance improvement in online adaptation—e.g., increasing success rates from 0.30 to 0.90 within minutes in real-world tasks—highlighting its robustness and versatility across training regimes.

Our main contributions are threefold. First, we propose a novel end-to-end goal-conditioned policy that integrates a visual world model with motor control, enabling robust zero-shot generalization across unseen objects, environments, and goals. Second, we introduce MSTH, a temporal abstraction mechanism that generates proximal and distal frames to balance long-horizon planning with closed-loop local control. Third, we develop an efficient online adaptation strategy that combines HER-style relabeling with LoRA-based fine-tuning, allowing for fast autonomous improvement on out-of-distribution scenarios. Together, these contributions establish a general and scalable framework for long-horizon goal-conditioned manipulation.

II. Related Works

Our Act2Goal method combines a goal-conditioned world model with an action expert capable of online autonomous improvement. Accordingly, we review related work in goal-conditioned policies, world models for robotic control, and online autonomous improvement.

A. Goal-conditioned Policy

Goal-conditioned policy learning aims to train agents to achieve diverse forms of goals, such as visual observations [2, 3, 4, 16, 17, 18, 19], spatial points [20, 21], or motion fields [22]. While these formulations vary in structure, visual goals remain the most expressive and widely applicable, especially for real-world manipulation tasks. Recent advances in this line of work have further explored long-horizon reasoning, keyframe-based planning [23], and subgoal generation [5, 24], to better accommodate the long-horizon requirements commonly encountered in real-world settings.

GoalGAIL [3] incorporates goal conditioning into imitation learning and accelerates policy convergence. However, long-horizon tasks remain challenging, as action-level imitation alone struggles to ensure consistent progress toward distant goals. To address this, GO-DICE [16] segments tasks into sub-goals, and CoA [23] generates reverse action sequences from goal keyframes. ManualVLA [24] also uses visual goals, but relies on external language prompts and handcrafted sub-goal generation.

These methods typically struggle to align current observations with distant goals. Act2Goal addresses these limitations by introducing a Goal-Conditioned World Model (GCWM) to simulate structured visual trajectories and proposing Multi-Scale Temporal Hashing to support consistent and efficient long-horizon planning, enabling better generalization in unseen tasks.

B. World Model for Robotic Control

World models have become a powerful tool in robotic control, enabling agents to simulate environmental dynamics [25, 26, 27], generate synthetic data for training [28, 29], or serve as learned simulators to guide policy learning [30, 31]. Recent works further combine world models with action experts (AEs) to form policy planning systems [32, 33, 12, 11], where the world model provides future state feature and the AE predicts actions accordingly.

Among these, GE-Act [13] adopts a bi-model architecture with a world model predicting future visual feature based on language instruction and a transformer-based planner generating actions. WorldVLA [34] jointly predicts vision and action in a unified latent space, aiming for tighter vision-action alignment.

Different from prior works, Act2Goal leverages a purely vision-based GCWM to guide policy learning with structured visual trajectories. To our knowledge, no prior work has integrated a world model into goal-conditioned policy learning.

C. Online Autonomous Improvement

To enhance policy adaptability during deployment, recent studies explore online improvement via either interactive imitation learning such as DAgger [35, 36, 37] or In-Context Learning (ICL) [38, 39]. However, DAgger-style methods require frequent expert intervention, while ICL methods do not update model weights and often show limited performance on complex tasks.

An alternative line of work leverages Hindsight Experience Replay (HER) [14] and its extensions [40, 41, 42], which relabel transitions by replacing original goals with achieved states, then optimize the policy using reinforcement or imitation learning. While these methods reduce the need for explicit rewards, they still rely on complex reward relabeling or external reward signals.

Act2Goal takes a further step toward fully autonomous online improvement. By combining HER-style relabeling with efficient LoRA-based finetuning, it enables direct policy adaptation from self-collected rollouts, without any task rewards or human annotations. This yields a lightweight, fully self-supervised update mechanism suitable for real-world deployment.

III. From World Model To General Goal-conditioned Policy

Our framework is designed to address two key challenges in long-horizon goal-conditioned manipulation. First, how to align low-level action generation with the desired scene configurations implied by a target visual goal. Second, how to balance long-term goal awareness with local reactivity—i.e., maintaining consistent progress toward the goal while robustly

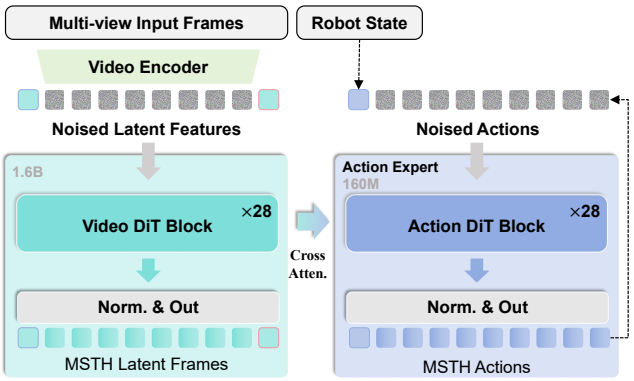

Fig. 3: **Model Architecture.** This figure presents the network architecture of Act2Goal model. On the left, multi-view input frames, including current observation and goal, are encoded into latents via a video encoder and concatenated with noisy latents, then refined into MSTH latent frames through Video DiT blocks. On the right, the robot state and multi-scale features from the world model are fed via cross-attention into isomorphic Action DiT blocks, generating MSTH-structured actions.

responding to perturbations during closed-loop execution. As illustrated in Figure 2, we address these challenges with two core components: a Goal-Conditioned World Model (GCWM) that generates imagined visual futures to provide temporally grounded guidance, and a temporal abstraction mechanism termed Multi-Scale Temporal Hashing (MSTH) that enables planning over both short-term execution and long-term intent.

Our learning process consists of three stages. Stage 1 jointly trains the GCWM and action expert to align perception and control through shared latent representations. Stage 2 focuses on action adaptation, further improving the policy's performance. Stage 3 introduces optional autonomous improvement, allowing the model to self-adapt in novel scenarios during deployment. We describe each component in detail below.

A. Goal-Conditioned World Model-Guided Policy

As illustrated in Fig. 3, the proposed GCWM builds upon the Genie Envisioner architecture with key modifications tailored for goal-conditioned policy learning [13]. We introduce a goal visual condition that is concatenated with the current observation along the hidden states sequence, while removing all language-conditioning components to create a purely vision-based model.

The GCWM leverages a continuous flow matching framework [43] to generate latent visual trajectories conditioned on the current and goal observations. Instead of directly mapping noise to structured sequences, the model learns a time-dependent vector field v_θ that transports a noisy latent variable ϵ toward the target trajectory over multiple steps.

Formally, given the VAE-compressed latents o_i and o_g corresponding to the current and goal observations, we initialize the process with $z^{(0)} = \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, and iteratively refine the prediction as:

$$z^{(n+1)} = z^{(n)} + \frac{1}{N} v_\theta(t_n, z^{(n)}, o_i, o_g), \tag{1}$$

where $t_n = 1 - \frac{n}{N}$ denotes the normalized time step. The vector field v_θ is trained to align this flow process with the true data distribution using a flow matching loss during offline training. After N steps, the predicted latent frames $z^{(N)}$ are decoded into visual observations using the VAE decoder.

Following the implementation of GE-Act [13], our action expert adopts a DiT-based architecture that is isomorphic to the world model, sharing the same number of transformer blocks but with reduced channel width. The action trajectory is generated via a flow matching process conditioned on both the robot's proprioceptive state c_p and the multi-scale visual features $c_w = \{h^1_{\text{world}}, \ldots, h^L_{\text{world}}\}$ produced by the world model.

At inference time, actions are generated by iteratively refining a noise vector through the learned vector field u_ϕ, following:

$$a^{(n+1)} = a^{(n)} + \frac{1}{N} u_\phi(t_n, a^{(n)}, c_w, c_p), \tag{2}$$

where $t_n = 1 - \frac{n}{N}$ is the normalized time step.

B. Multi-Scale Temporal Hashing for Visual State and Action

In our framework, the MSTH mechanism jointly guides the goal-conditioned world model and the action expert, enabling a consistent yet flexible multi-scale temporal abstraction. Given a total imagined trajectory length K, a proximal horizon P, and a vision sampling stride r, MSTH partitions the future trajectory into two segments.

The proximal segment consists of high-frequency short-horizon visual states $\{s_{t+kr}\}_{k=1}^{P/r}$, which capture fine-grained local dynamics. The distal segment contains M sparsely sampled visual states $\{s_{t+d_m}\}_{m=1}^{M}$, where the indices d_m are determined by logarithmic spacing:

$$d_m = P + \left\lfloor \frac{K - P}{\log(M+1)} \cdot \log(m+1) \right\rfloor. \tag{3}$$

This logarithmic sampling results in increasing temporal intervals as the horizon extends, providing coarse but goal-aligned long-term guidance.

The predicted action sequence follows the same multi-scale structure, with an important distinction from vision. Proximal actions are predicted at every timestep, $\{a_{t+1}, a_{t+2}, \ldots, a_{t+P}\}$, enabling dense motor control even when visual states are subsampled by stride r. In contrast, distal actions $\{a_{t+d_m}\}_{m=1}^{M}$ are aligned with the distal visual states and serve as long-horizon guidance. During deployment, only the proximal actions are executed, while distal predictions act as latent guidance for long-horizon goal adherence.

C. Two-Stage Offline Training

To endow Act2Goal with strong generalization capabilities, we first train the model through large-scale offline imitation learning. The training process consists of two main stages. The first stage focuses on jointly aligning the visual trajectory

prediction task with action planning, while the second stage refines the policy to better match expert demonstrations in an end-to-end manner.

In the first stage, we fine-tune a pre-trained visual world model for the task of generating goal-conditioned video trajectories that follow the MSTH distribution between the initial observation and target goal. To encourage consistency between the predicted visual transitions and the downstream action policy, we jointly train the world model and the action head using flow matching objectives.

The visual and action components are optimized using the following flow matching losses:

$$\mathcal{L}_v = \mathbb{E}\left[\|v_\theta(t, z_t, o_i, o_g) - (z_1 - z_0)\|^2\right] \tag{4}$$

$$\mathcal{L}_a = \mathbb{E}\left[\|u_\phi(t, a_t, c_w, c_p) - (a_1 - a_0)\|^2\right]. \tag{5}$$

Here, z_0 and a_0 are sampled from demonstration trajectories, while z_1 and a_1 are sampled from standard Gaussian noise. The interpolated inputs $z_t = t \cdot z_1 + (1-t) \cdot z_0$ follow a linear schedule with $t \sim \mathcal{U}(0, 1)$.

The total loss for this stage is a weighted sum of the visual and action components:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_v + \lambda \cdot \mathcal{L}_a, \tag{6}$$

where $\lambda = 0.1$ is a balancing coefficient. This joint optimization encourages the world model to generate visual trajectories that are both temporally coherent and informative for action prediction.

In the second training stage, we employ behavioral cloning to fine-tune the entire model end-to-end using only the action flow matching loss $\mathcal{L}_{\text{stage2}} = \mathcal{L}_a$, further strengthening its action planning capabilities. This stage focuses on aligning the complete pipeline—from visual perception to action execution—with expert demonstrations. The gradients from the action loss propagate through both the action generation components and the goal-conditioned world model, allowing the visual representations to be optimized specifically for action planning.

Together, this two-stage offline training scheme equips Act2Goal with visual and motor priors that generalize effectively to novel manipulation scenarios.

D. Online Autonomous Improvement

While the model exhibits strong generalization capabilities after offline imitation learning, achieving high performance on novel tasks, environments, objects, and motion control chains remains challenging when deployed on physical robots—a common limitation of imitation learning-based policies. To address this, we introduce an online autonomous improvement framework based on Hindsight Experience Replay (HER), enabling autonomous performance enhancement during deployment [14]. This stage assumes that the offline policy already exhibits partial competence or at least meaningful interaction with the target object, so that its collected rollouts can provide useful relabeling signals for HER-based adaptation.

Algorithm 1 Act2Goal Online Autonomous Improvement

```
Initialize replay buffer B and LoRA parameters φ
while performance not converged do
    Roll out policy π_θ with goal g, collect (o, g, c_p, a, o') at each step
    Add collected transitions to buffer B
    if |B| ≥ N then
        for k = 1 to K do
            Sample batch {(o, g, c_p, a, o')} ~ B
            Relabel g' ← o'
            Compute loss using (o, g', c_p, a) and update φ
        end for
        Clear buffer: B ← ∅
    end if
end while
```

As illustrated in Algorithm 1, the system operates as follows. Given a target goal condition g, the model infers and executes an action a based on the current observation o and proprioceptive state c_p, resulting in a new observation o'. The transition tuple (o, g, c_p, a, o') is then stored in a replay buffer $\mathcal{B}$.

Once the buffer accumulates N transitions, we perform online policy updates for K steps. In each step, a batch of transitions is sampled from $\mathcal{B}$, and the final observation o' is relabeled as a new goal g', following the HER strategy. The associated action a is interpreted as the expert action for reaching g' from the corresponding observation o. The policy is then updated using the same flow matching objective as in Stage 2 of offline training. To enable efficient on-device learning, we update only the lightweight LoRA parameters introduced in the Act2Goal model, while keeping the base model frozen [15].

After each update cycle, the system resumes data collection through continued rollouts, repeating the loop until task performance converges. For safety and stability, we define a maximum number of inference steps per task; exceeding this threshold triggers an automatic reset of the robot. During real-world experiments, the only required human intervention is occasional manual resetting or adjustment of the environment when necessary.

IV. Experiments

Our experiments aim to systematically evaluate the effectiveness of the proposed Act2Goal model across both offline and online settings. We assess the model's ability to generalize to novel visual goals, objects, spatial arrangements, and task configurations, and to autonomously improve itself through interactions during deployment. Specifically, we study three key aspects: (1) the generalization capability of policies trained offline using imitation learning, measured under both in-domain (ID) and out-of-domain (OOD) scenarios; (2) the model's capacity for online autonomous improvement, enabled

TABLE I: **Comparison of Performance on Robotwin 2.0 Simulation Benchmark.** Act2Goal outperforms baselines in all Easy-mode tasks and 3 Hard-mode tasks, showing superior generalization.

	Model/Task	Move Can	Pick Bottles	Place Cup	Place Shoe
Easy	DP-GC	0.18	0.04	0.03	0.04
	HyperGoalNet	0.11	0.08	0.08	0.01
	$\pi_{0.5}$-GC	0.54	0.13	0.16	0.30
	Act2Goal	**0.62**	**0.80**	**0.64**	**0.52**
Hard	DP-GC	0.00	0.00	0.00	0.00
	HyperGoalNet	0.00	0.00	0.00	0.00
	$\pi_{0.5}$-GC	**0.42**	0.06	0.04	0.06
	Act2Goal	0.13	**0.43**	**0.13**	**0.15**

by lightweight on-device fine-tuning of LoRA layers with self-relabelled transitions; and (3) the contribution of our proposed Multi-Scale Temporal Hashing (MSTH) mechanism, evaluated through qualitative analysis and ablations on both ID and OOD tasks.

A. Generalization Capability Evaluation

To thoroughly evaluate the generalization ability and versatility of Act2Goal after offline imitation training, we conduct extensive experiments in both simulation and real-world robotic settings, under both ID and OOD configurations. In this work, OOD primarily refers to goal-level generalization. Specifically, the target visual states, object instances, spatial arrangements, or task configurations differ from those observed during training. Our evaluation is therefore designed to test whether the policy can reach substantially different goal states from visual specifications.

We compare Act2Goal with the following representative goal-conditioned policy baselines, and all baselines are trained on task-specific data with GC objective unless stated otherwise:

- DP-GC: A DiT-style action policy [44] that uses cross-attention to fuse SigLIP features [45] of the current observation and goal image.
- HyperGoalNet [19]: A recent high-performance open-source goal-conditioned policy architecture.
- $\pi_{0.5}$-GC: A vision-language-action model that uses fixed language descriptions but incorporates both the goal image and current observation as input; it uses the original $\pi_{0.5}$ [46] weights.
- $\pi_{0.5}$-GC (pretrained): An enhanced version that loads the same $\pi_{0.5}$ checkpoint but is pretrained on our GC pretraining data, using the GC objective.

We first evaluate these methods on the Robotwin 2.0 simulation benchmark [47], selecting four manipulation tasks with both Easy and Hard variants. For each test episode, the goal image is extracted from a successful trajectory under a fixed seed and used as the conditioning input. This benchmark provides a controlled setup for assessing generalization ability.

As reported in Table I, Act2Goal achieves notably high success rates in Easy scenarios, while also maintaining robust performance under the more challenging Hard conditions—highlighting its strong generalization ability across a wide range of visual and spatial variations.

TABLE II: **Comparison of Performance on Real-World Manipulation Tasks.** Act2Goal demonstrates strong performance across all real-world tasks, significantly outperforming all baseline methods. The results support strong goal-level generalization under the evaluated OOD settings. All results are obtained without any online autonomous improvement, using only offline imitation learning.

	Model/Task	Whiteboard Word Writing	Dessert Plating	Plug-In Operation
ID	DP-GC	0.00	0.10	0.00
	HyperGoalNet	0.00	0.08	0.00
	$\pi_{0.5}$-GC	0.23	0.18	0.00
	$\pi_{0.5}$-GC (pretrained)	0.58	0.68	0.25
	Act2Goal	**0.93**	**0.75**	**0.45**
OOD	DP-GC	0.00	0.00	0.00
	HyperGoalNet	0.00	0.00	0.00
	$\pi_{0.5}$-GC	0.20	0.05	0.00
	$\pi_{0.5}$-GC (pretrained)	0.30	0.45	0.08
	Act2Goal	**0.90**	**0.48**	**0.30**

To assess real-world applicability, we deploy the model on an AgiBot Genie-01 robot across three challenging manipulation tasks. These tasks are difficult to specify using language and instead require goal image conditioning. Importantly, the test tasks are drawn from the offline training dataset but do not involve any task-specific supervised fine-tuning—thus directly testing the generalization capacity of Act2Goal.

Each task is evaluated in both ID and OOD variants using success rate as the metric:

- **Task 1: Whiteboard Word Writing.** The robot writes English words. ID words were seen during training (200 words total); OOD words feature unseen combinations, testing compositional generalization.
- **Task 2: Dessert Plating.** The robot arranges desserts on a plate. ID uses dessert types, plates, and backgrounds seen during training; OOD introduces novel styles and strong distractors to test visual robustness.
- **Task 3: Plug-In Operation.** The robot inserts objects. ID involves inserting a seen metal workpiece; OOD replaces this with inserting a cylindrical drink bottle—an unseen task testing physical skill transfer.

As summarized in Fig. 4 and Table II, Act2Goal consistently achieves high success rates across both ID and out-of-domain OOD tasks, significantly outperforming all competing baselines. Its strong performance demonstrates the capacity of our approach to generalize to novel objects, spatial configurations, and goal states under the evaluated OOD settings. These results support the practicality of Act2Goal for diverse real-world goal-conditioned manipulation scenarios.

B. Analysis of Online Autonomous Improvement

We evaluate the effectiveness of online autonomous improvement in both simulation and real-world settings, and systematically analyze strategies for enhancing robot manipulation performance through Hindsight Experience Replay (HER). These experiments focus on scenarios where the initial

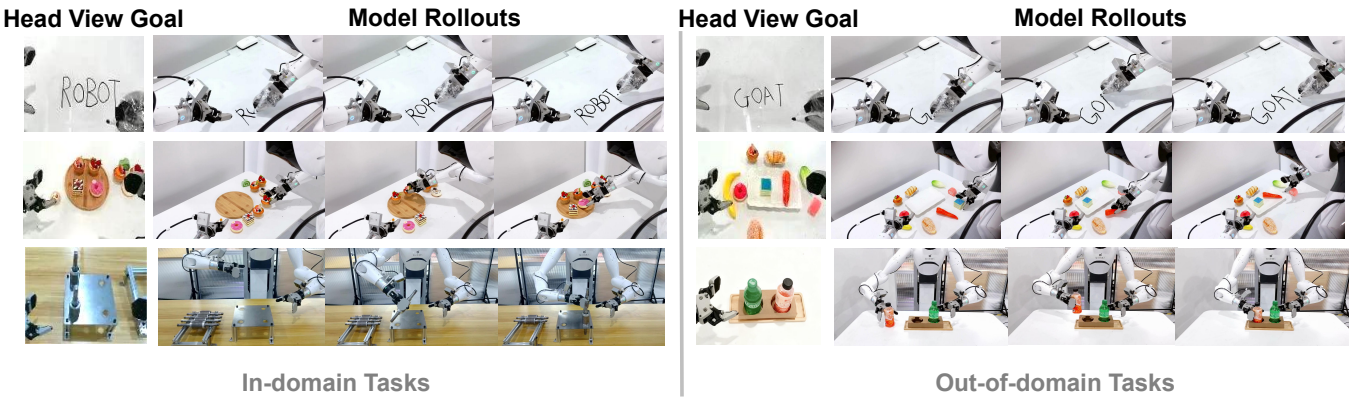

Fig. 4: **Real World Evaluation.** This figure illustrates in-domain and out-of-domain test configurations for three real-world tasks: Whiteboard Word Writing, Dessert Plating, and Plug-In Operation. For each task, Head View Goal displays the target, while Model Rollouts shows the robot's execution process; these setups are used to evaluate the model's generalization ability using the success rate as the metric.

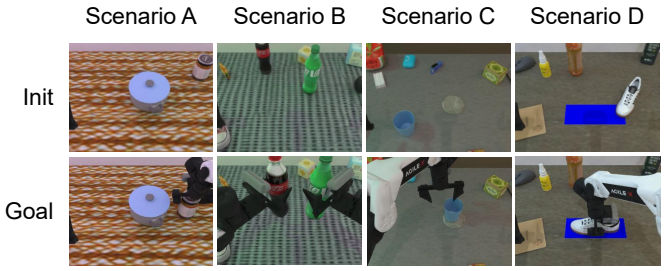

Fig. 5: **Online Autonomous Improvement Scenarios.** This figure illustrates four OOD scenarios from the RoboTwin 2.0 benchmark, corresponding to the hard testing modes of Move Can Pot, Pick Dual Bottles, Place Empty Cup, and Place Shoe. These scenarios serve as the testbed for verifying the effectiveness of autonomous improvement.

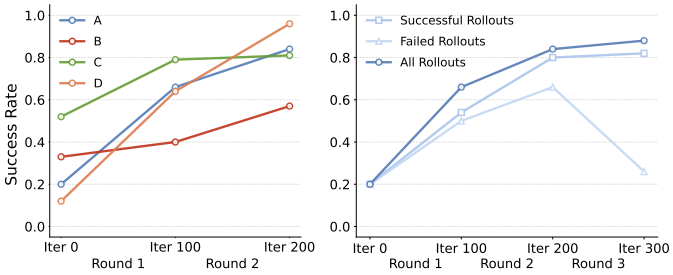

Fig. 6: **Online Training Performance in Robotwin 2.0.** This figure presents two key findings from Robotwin 2.0 simulations: (Left) Multi-round success rates of four hard-mode scenarios, showing consistent improvement over 3 rounds before convergence. (Right) Performance of three data selection strategies for rollouts: using all rollouts yields optimal results, while even failed-only rollouts enable noticeable improvement.

policy already produces meaningful interactions, rather than learning entirely new skills from scratch.

In the Robotwin 2.0 simulator, we selected one Hard mode

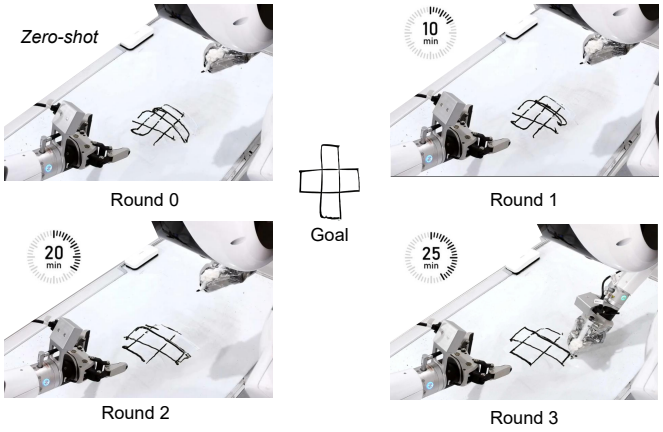

Fig. 7: **Online Training Performance in Real-World Unseen Scenarios.** When tasked with drawing an unseen pattern, the model initially performs poorly. However, as online training progresses (from left to right, top to bottom), the drawing quality improves steadily.

configuration from each of the four tasks described in Section IV-A (shown in Fig. 5) and performed multiple rounds of online training. After each round, we measured the success rate to track performance changes. As shown in the left part of Fig. 6, Act2Goal consistently improves across all scenarios and typically converges after three rounds, with up to an 8× success rate increase over the pre-trained baseline.

We further investigated different data selection strategies for populating the online replay buffer: (1) using only successful rollouts, (2) using all rollouts regardless of outcome, and (3) using only failed rollouts. As illustrated in the right part of Fig. 6, training on all rollouts leads to the highest final performance. Interestingly, using only failed rollouts also yields substantial improvement, confirming that HER can effectively extract learning signals even from failure cases—a key benefit in goal-conditioned policy refinement.

In real-world experiments, we conducted online adaptation

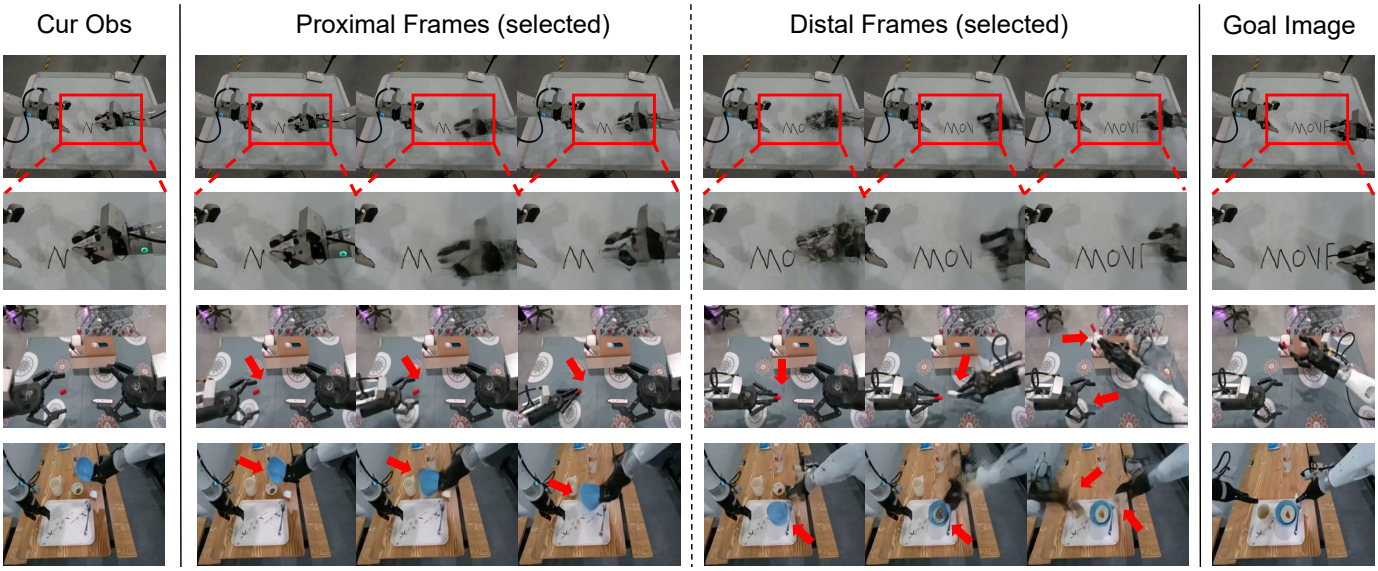

Fig. 8: **Examples of generated videos.** This figure illustrates the generation capability of our goal-conditioned world model through three head-view video clips. The left and right sides show the current observation and the target goal, respectively. For each sequence, we uniformly sample three proximal and three distal frames. The second row provides a zoomed-in view of the red-highlighted regions in the first row. Red arrows in the third and fourth rows highlight the motion of key objects. The distal frames in the bottom row correspond to longer temporal horizons compared to those in the third row.

TABLE III: **Effectiveness of MSTH.** Values denote the success rate of the real-world whiteboard word writing task. The word length definitions: Short (3 letters or fewer), Medium (4 to 6 letters), Long (7 letters or more).

	Model	Short	Medium	Long
ID	w/o MSTH	0.95	0.35	0.10
	w/ MSTH	**0.95**	**0.90**	**0.90**
OOD	w/o MSTH	0.60	0.20	0.00
	w/ MSTH	**0.93**	**0.90**	**0.88**

on two tasks refered in Section IV-A. In the whiteboard writing task, While the model exhibits strong generalization in whiteboard word writing, we assessed its capability to draw novel patterns that are substantially different from writing.As shown in Fig. 7, the model's initial performance on this unfamiliar case was limited. Nevertheless, through online fine-tuning, it demonstrated consistent improvements in drawing quality in 25 minutes. For the Plug-In task, the model likewise exhibited steady performance gains during online rollout, with success rate improving from 0.30 to 0.90 over time. Details are in the supplementary video material provided in the submission.

C. *Effectiveness of MSTH*

To evaluate the impact of Multi-Scale Temporal Hashing (MSTH), we analyze both qualitative and quantitative results of Act2Goal.

Fig. 8 showcases representative examples of video rollouts generated by our goal-conditioned world model. Each rollout illustrates both fine-grained proximal frames—providing detailed, short-horizon guidance crucial for generating precise and executable actions—and distal frames that capture high-level long-horizon structure. This multi-scale trajectory prediction enables the model to reason across temporal scales, offering both low-level motor control precision and high-level goal adherence. Such hierarchical visual forecasting serves as an effective internal plan, bridging the gap between immediate actions and the final goal.

To quantitatively assess the effectiveness of MSTH, we perform an ablation study on the real-world whiteboard word writing task, which inherently requires long-horizon, visually grounded planning. We compare the performance of our full method (with MSTH) against a baseline that employs fixed-horizon action chunking. As shown in Table III, we report success rates under both ID and OOD settings, stratified by word length: short (at most 3 letters), medium (4–6 letters), and long (at least 7 letters).

While both methods perform comparably on short words, the fixed-horizon baseline suffers significant performance degradation as word length increases, especially in the OOD setting. In contrast, the MSTH-based policy maintains consistently high success rates across all word lengths and generalizes well to unseen OOD words. This demonstrates MSTH's ability to dynamically adjust the level of temporal abstraction and maintain goal-aligned plans throughout extended executions. By introducing temporally adaptive anchors and mitigating goal drift, MSTH proves to be a simple yet powerful mechanism for enhancing long-horizon goal-conditioned behavior.

TABLE IV: **Additional comparison of temporal abstractions.** Values denote success rates on the OOD fruit plating task. All models are trained with 1000 pick-place trajectories under a shorter-horizon setting with maximum observation-goal distance 384.

Model	Exec. 30	Exec. 50	Exec. 120
Fixed-length	0.13	0.15	–
Dense	0.00	0.05	0.60
MSTH (log)	**0.60**	**0.73**	–
MSTH (linear)	0.55	0.70	–

To further isolate the effect of temporal abstraction, we conduct an additional controlled comparison among fixed-length, dense, and MSTH abstractions in a shorter-horizon setting where dense supervision is computationally feasible. Specifically, we set the maximum observation-goal distance to 384, train all models using 1000 pick-place trajectories, and evaluate them on the OOD fruit plating task. As shown in Table IV, the fixed-length abstraction performs poorly because it only predicts the first few future states and can therefore provide weak goal-specific supervision when different goals share similar early trajectories. The dense abstraction also performs poorly under realistic execution horizons, but becomes competitive when using an impractically long execution chunk. In contrast, MSTH achieves strong performance under realistic horizons by combining dense proximal guidance with sparse distal goal anchors. Moreover, the small gap between logarithmic and linear distal sampling suggests that the main benefit comes from the multi-scale temporal abstraction itself, rather than from a specific hand-crafted sampling rule.

Together, the real-world ablation and the controlled temporal-abstraction comparison highlight MSTH as a key component for scaling goal-conditioned policies to complex, temporally extended manipulation tasks.

V. Conclusion

We propose Act2Goal, a goal-conditioned policy that combines a visual world model with multi-scale temporal control to tackle long-horizon manipulation. The model generates both fine-grained short-term executable control signals and coarse long-term trajectories, enabling effective planning and control. Trained fully offline, Act2Goal achieves high success rates in seen tasks and shows strong generalization to unseen scenarios. To enhance adaptability, we incorporate Hindsight Experience Replay and LoRA-based finetuning, enabling efficient, reward-free online autonomous improvement within minutes. Together, these components contribute to a scalable and generalizable visuomotor policy for open-ended, real-world manipulation tasks.

VI. Limitations

A main limitation of the current online autonomous improvement stage is its cold-start requirement. The HER-based adaptation assumes that the initial policy already has partial competence, or at least produces meaningful interactions with the target object, so that its rollouts can provide useful relabeling signals. It is therefore not intended to acquire entirely novel skills from scratch when the required behavior is absent from the training distribution, such as zipper closing without related prior experience. Addressing this limitation may require broader pretraining data, additional demonstrations, or external guidance.

References

[1] Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, volume 2, pages 1094–8, 1993.

[2] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022.

[3] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. *Advances in neural information processing systems*, 32, 2019.

[4] Xudong Gong, Dawei Feng, Kele Xu, Bo Ding, and Huaimin Wang. Goal-conditioned on-policy reinforcement learning. *Advances in neural information processing systems*, 37:45975–46001, 2024.

[5] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

[6] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024.

[7] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. *arXiv preprint arXiv:2304.02532*, 2023.

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[9] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[10] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[11] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[12] Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37:41051–41075, 2024.

[13] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

[14] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[16] Abhinav Jain and Vaibhav Unhelkar. Go-dice: Goal-conditioned option-aware offline imitation learning via stationary distribution correction estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12763–12772, 2024.

[17] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation. *IEEE Transactions on Robotics*, 40:2287–2305, 2024.

[18] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[19] Pei Zhou, Wanting Yao, Qian Luo, Xunzhe Zhou, and Yanchao Yang. Hyper-goalnet: Goal-conditioned manipulation policy learning with hypernetworks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[20] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[21] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pages 306–324. Springer, 2024.

[22] Zhao-Heng Yin, Sherry Yang, and Pieter Abbeel. Object-centric 3d motion field for robot learning from human videos. *arXiv preprint arXiv:2506.04227*, 2025.

[23] Wenbo Zhang, Tianrun Hu, Yanyuan Qiao, Hanbo Zhang, Yuchu Qin, Yang Li, Jiajun Liu, Tao Kong, Lingqiao Liu, and Xiao Ma. Chain-of-action: Trajectory autoregressive modeling for robotic manipulation. *arXiv preprint arXiv:2506.09990*, 2025.

[24] Chenyang Gu, Jiaming Liu, Hao Chen, Runzhong Huang, Qingpo Wuwu, Zhuoyang Liu, Xiaoqi Li, Ying Li, Renrui Zhang, Peng Jia, Pheng-Ann Heng, and Shanghang Zhang. Manualvla: A unified vla model for chain-of-thought manual generation and robotic manipulation, 2025. URL https://arxiv.org/abs/2512.02013.

[25] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

[26] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[27] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[28] Haoyu Zhao, Cheng Zeng, Linghao Zhuang, Yaxi Zhao, Shengke Xue, Hao Wang, Xingyue Zhao, Zhongyu Li, Kehan Li, Siteng Huang, et al. High-fidelity simulated data generation for real-world zero-shot robotic manipulation learning with gaussian splatting. *arXiv preprint arXiv:2510.10637*, 2025.

[29] Liu Liu, Xiaofeng Wang, Guosheng Zhao, Keyu Li, Wenkang Qin, Jiaxiong Qiu, Zheng Zhu, Guan Huang, and Zhizhong Su. Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer. *arXiv preprint arXiv:2505.23171*, 2025.

[30] Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

[31] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

[32] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

[33] Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

[34] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[35] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[36] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Inter-

active imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083. IEEE, 2019.

[37] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10 (105):eads5033, 2025.

[38] Rutav Shah, Shuijing Liu, Qi Wang, Zhenyu Jiang, Sateesh Kumar, Mingyo Seo, Roberto Martín-Martín, and Yuke Zhu. Mimicdroid: In-context learning for humanoid robot manipulation from human play videos. *arXiv preprint arXiv:2509.09769*, 2025.

[39] Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. Ricl: Adding in-context adaptability to pre-trained vision-language-action models. *arXiv preprint arXiv:2508.02062*, 2025.

[40] Rui Yang, Meng Fang, Lei Han, Yali Du, Feng Luo, and Xiu Li. Mher: Model-based hindsight experience replay. *arXiv preprint arXiv:2107.00306*, 2021.

[41] Liam Schramm, Yunfu Deng, Edgar Granados, and Abdeslam Boularias. Usher: Unbiased sampling for hindsight experience replay. In *Conference on Robot Learning*, pages 2073–2082. PMLR, 2023.

[42] Yongle Luo, Yuxin Wang, Kun Dong, Qiang Zhang, Erkang Cheng, Zhiyong Sun, and Bo Song. Relay hindsight experience replay: Self-guided continual reinforcement learning for sequential object manipulation tasks with sparse rewards. *Neurocomputing*, 557:126620, 2023.

[43] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[44] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

[45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[46] 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. URL https://arxiv.org/abs/2504.16054.

[47] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.