2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

From Dream to Action: Hierarchical Policy Learning with 3D World Imagination for Robotic Manipulation

Wenshuo Wang[1], Ruiteng Zhao[1], Tat Joo Teo[2], Marcelo H. Ang Jr.[3] and Haiyue Zhu[4†]

***Abstract*— Recent advancements in robotics have focused on developing foundation models capable of generating both actions and future states. Typically, these policies leverage world models to depict human-like imagination. However, most methods remain confined to the 2D domain, where they forecast only the final outcome state rather than the evolving interaction process, thereby offering limited guidance for step-by-step control. To address these limitations, we propose a hierarchical framework that couples 3D imagination, 3D perception, and action generation. A triplane-based world model captures future scene dynamics in a computationally efficient manner, providing predictive cues for decision-making. Based on these representations, the action expert, implemented with a flow-based policy network, converts the outputs of 3D imagination and perception into executable commands. We further introduce an adaptive Classifier-Free Guidance strategy to balance action quality with condition adherence. On Adroit, Meta-World, and real-world tasks, our method achieves a 92% voxel IoU in future state prediction and up to 8% higher success rates than state-of-the-art baselines. The performance gains highlight the effectiveness and generalizability of our method in complex robotic manipulation.**

I. Introduction

Imitation learning has emerged as a prominent research direction for robot manipulation [1]–[3], which enables robots to acquire complex skills from expert demonstrations. It provides a powerful paradigm for mapping high-dimensional visual observations to low-level motor commands. In recent years, advances in visuomotor policy learning have been driven by large-scale embodied foundation models [4]–[6]. Leveraging open-source embodied datasets [7]–[9], such models are capable of performing both high-level planning and low-level control within the 2D image space. Nevertheless, 2D observations inherently discard rich geometric and spatial cues. For robust and generalizable skill learning, it is also essential for robots to exploit the spatial information in 3D modalities.

This research is supported by National Robotics Programme (NRP) 2.0 funding initiative "Domain-specific Robotics Foundation Models for Manufacturing (DS-RFM)".

[1]Wenshuo Wang and Ruiteng Zhao are with the Advanced Robotics Centre, National University of Singapore, and are also attached students of SIMTech, A*STAR. (email: {wenshuo_wang, ruiteng}@u.nus.edu)

[2]Tat Joo Teo is with the Robotics, Automation & Unmanned Systems Centre of Expertise, Home Team Science and Technology Agency (HTX), Singapore 138507 (email: daniel_teo@htx.gov.sg)

[3]Marcelo H. Ang Jr is with Advanced Robotics Centre at National University of Singapore, Singapore 117608, (email: mpeangh@nus.edu.sg)

[4]Haiyue Zhu is with the Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore 138634, (email: zhu_haiyue@simtech.a-star.edu.sg)

† Corresponding author: zhu_haiyue@simtech.a-star.edu.sg

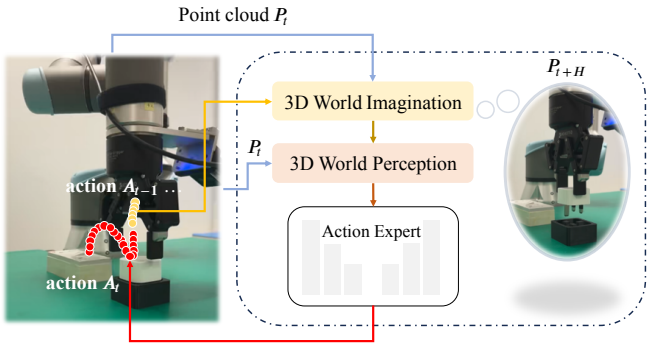

Fig. 1: Our hierarchical policy leverages 3D world imagination and 3D world perception to anticipate intermediate dynamics and guide sequential action generation.

Taking one step forward, recent works [10]–[12] have explored visuomotor policies conditioned on 3D representations, thereby mitigating the loss of spatial information inherent in 2D methods. Despite this advantage, these methods still map observations directly to actions, without anticipating interactive dynamics within the 3D scene space. We refer to this limitation as the "Myopic Phenomenon". Inspired by human cognition [13], where individuals leverage an internal world model to mentally simulate the evolution of the environment, we aim to explore whether, and in what manner, the imagination of 3D future states can enhance robots' ability to generate robust and effective action trajectories.

Another challenge in building a 3D world model-based policy lies in balancing efficiency and quality. On one hand, injecting anticipatory ability into embodied models inevitably incurs additional computational overhead during both the training and inference processes. On the other hand, real-time decoding of high-quality robot actions often suffers from runtime inefficiencies. Therefore, designing a framework that not only reasons about future states but also maintains a lightweight architecture is essential for scalable and responsive robotic manipulation.

To this end, we propose a compact and efficient framework that seamlessly synergizes 3D world imagination, 3D world perception, and action generation. Instead of following the end-to-end WorldVLA [14] or 3D-VLA [15] models, we develop a hierarchical architecture that decouples imagination from decision-making. As illustrated in Fig. 1, we first introduce a 3D world model that leverages the triplane representation [16] to simulate future states. The triplane representation can reduce the computation of unnecessary empty information through the factorization of 3D data into

2D planes [16]. Conditioned on the current point cloud and historical actions, the world model explicitly reconstructs the next intermediate future point cloud. Another challenge in building such an efficient framework lies in the design of action experts. Diffusion-based solutions typically require numerous sampling steps. Thus, we build upon the flow-based architecture as the backbone, and further introduce adaptive Classifier-Free Guidance to balance the trade-off between generation quality and condition adherence. Our framework bridges imagination and decision-making through a hierarchical design, offering both computational efficiency and adaptability to complex robotic tasks.

The experiments on Adroit and the Meta-World benchmark demonstrate that our method outperforms the state-of-the-art flow-based baseline by an average margin of 8% in success rate. The learned world model achieves 92% voxel IoU accuracy, suggesting that the proposed representation not only facilitates more accurate prediction of future states but also enhances downstream control. In real-world evaluations, our policy succeeds on long-horizon Cook and contact-rich tasks such as Wiping and Insertion, where prior baselines consistently fail. These results highlight the effectiveness and generalizability of our approach across a wide range of manipulation scenarios.

In summary, our contributions are as follows:

- We develop a hierarchical framework that unifies 3D imagination, perception, and action generation.
- The triplane-based world model serves as a plug-and-play module for diverse visuomotor policies.
- We introduce an adaptive Classifier-Free Guidance strategy to effectively balance the trade-off between action generation quality and adherence to visual conditions.
- We conduct extensive experiments on Adroit, Meta-World, and real-world tasks, demonstrating that our approach achieves state-of-the-art success rates.

II. Related Works

A. World Imagination

In pursuit of Artificial General Intelligence (AGI), large-scale multimodal models [17] have been developed to capture and model the dynamics of the external world. A representative example is Sora [18], which simulates future world evolution by generating high-quality video frames. Beyond video generation [19], [20], world models have also found applications in autonomous driving [21] and robotics [22].

In robotics, world models are typically leveraged to envision possible future states that guide the decision-making process. Particularly, RoboDreamer [22] proposes a compositional approach to capture spatial relationships and object interactions in robotic video generation. 3D-VLA [15] introduces a 3D embodied foundation model that connects perception, reasoning, and action through a generative world model. In addition, WorldVLA [14] develops an autoregressive action world model for unified action and image generation. However, these frameworks primarily predict only long-horizon final states. In contrast, our method explicitly predicts intermediate future states, enabling more efficient and fine-grained modeling of temporal evolution.

B. Action Experts

In visuomotor policies, action experts are designed to generate low-level actions from visual representations. Recently, diffusion models [23] define a forward process that perturbs data into noise, and then learn a reverse process to progressively denoise it. The denoising process is typically formulated as solving either a Stochastic Differential Equation (SDE) or an Ordinary Differential Equation (ODE). As a pioneer, Diffusion Policy [24] was among the first to leverage diffusion models for robotic action generation. Subsequently, several studies [12], [25] extended this approach to broader domains. However, diffusion-based methods typically require numerous sampling steps during inference, which are inevitably plagued by substantial runtime inefficiencies.

Flow matching [26] addresses this limitation by modeling the velocity field that drives the noise distribution toward the data distribution. Instead of relying on iterative denoising steps, flow matching provides a direct probabilistic path from noise to the target. In robotics, several works [27]–[29] have explored its variants for real-time action generation. FlowPolicy [28] leverages the conditional generation capability of Consistency Flow Matching [30] for embodied tasks. MP1 [29] alternatively explores the MeanFlow [31] paradigm for robot learning. In this work, we build upon the state-of-the-art MeanFlow method and further introduce an adaptive CFG strategy to balance the trade-off between sample diversity and conditional consistency. Without loss of generality, our world model is designed as a plug-and-play module that can be seamlessly incorporated into both diffusion- and flow-based policy frameworks. The framework achieves efficient one-step inference and generates high-quality robot actions.

III. Methodology

A. Motivation

In this section, we present the motivation behind our proposed framework. While most visuomotor policies directly map short-term past observations $o \in O$ to actions $a \in A$, the myopic formulation constrains their capacity to reason about interactive world dynamics. To achieve robust visuomotor control, an effective policy requires 1) a perception module that captures a temporally consistent, spatially aware representation of the environment, and 2) an efficient action expert capable of leveraging these predictions to produce goal-directed trajectories. These considerations motivate a hierarchical architecture: 1) **3D World Imagination.** The 3D world model is first trained to anticipate future visual states explicitly based on current observations and prior actions; 2) **3D World Perception.** The imagined and observed point clouds are encoded into compact 3D representations; 3) **Adaptive Action Generation.** A flow-based policy transforms noise samples into action trajectories with adaptive guidance to balance fidelity and condition

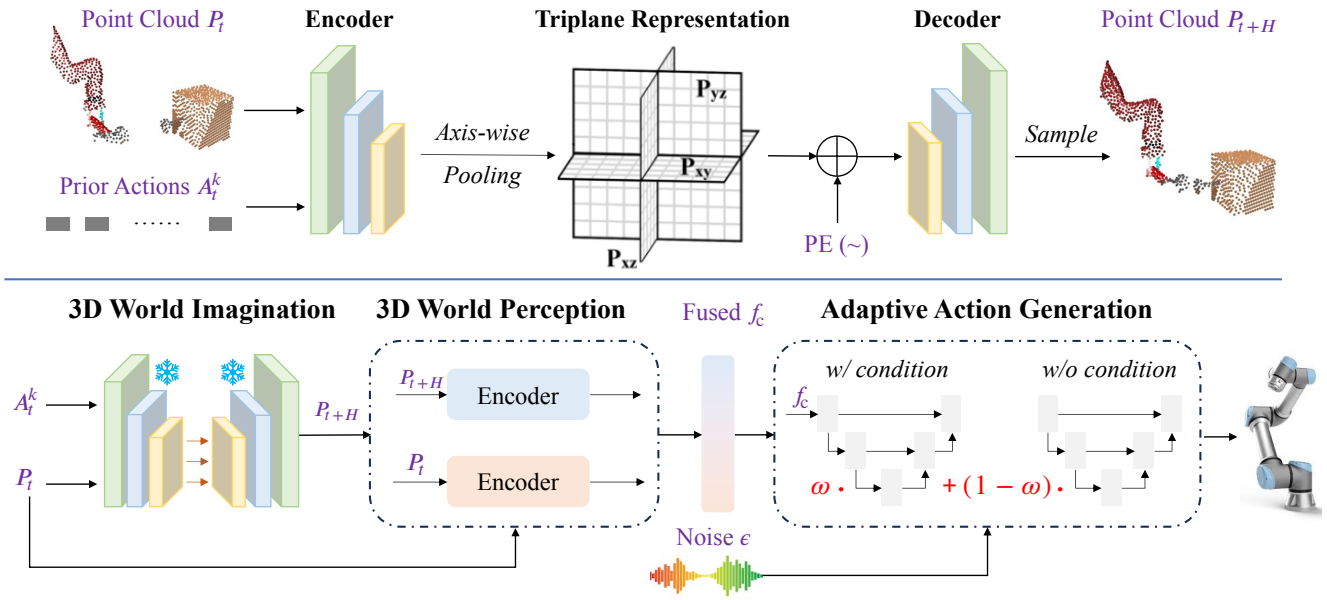

Fig. 2: Overview pipeline. (**Top**) **3D World Imagination.** The 3D world model is pre-trained to anticipate future visual states explicitly based on current observations and prior actions. (**Bottom**) **Action Generation.** The current point cloud, together with the imagined states from the learned 3D world model, is encoded and fused into compact 3D conditions. The flow-based action expert with adaptive CFG achieves high-quality, real-time action generation.

adherence. Fig. 2 illustrates the overall system architecture. The details of each stage are presented below.

B. 3D World Imagination

In the 3D World Imagination stage, we introduce two key components: 1) a triplane representation for scene modeling, and 2) a triplane autoencoder for scene reconstruction. As illustrated in Fig. 2 (Top), the pipeline begins by converting a 3D point cloud $\boldsymbol{P_t} \in \mathbb{R}^{N\times 3}$ into a voxel grid $\boldsymbol{V_t} \in \mathbb{R}^{H\times W\times D}$ with a predefined spatial resolution of H, W, and D. A convolutional encoder $\mathcal{E}$ is then leveraged to extract a 3D feature volume $\boldsymbol{F_t} \in \mathbb{R}^{C\times H'\times W'\times D'}$, where H', W', and D' denote the encoded spatial dimensions and C is the feature dimension. To achieve a more compact yet expressive scene representation, the feature volume is further factorized into three orthogonal 2D feature planes via axis-wise pooling, yielding $\boldsymbol{P_{xy}} \in \mathbb{R}^{C\times H'\times W'}$, $\boldsymbol{P_{xz}} \in \mathbb{R}^{C\times H'\times D'}$, and $\boldsymbol{P_{yz}} \in \mathbb{R}^{C\times W'\times D'}$. For a 3D point $\boldsymbol{p} \in \boldsymbol{P_t}$, its feature embedding $\boldsymbol{f_p}$ is obtained by aggregating the interpolated features from the three axis-aligned planes. Specifically, we compute

$$\boldsymbol{f_p} = \boldsymbol{P_{xy}}(x,y) + \boldsymbol{P_{xz}}(x,z) + \boldsymbol{P_{yz}}(y,z), \tag{1}$$

where x, y, and z denote the projection coordinates on the corresponding 2D planes. If (x,y,z) does not coincide with the discrete grid, we approximate its feature value by bilinear interpolation from the neighboring 2D grid values. $\boldsymbol{P_{xy}}(x,y)$ at a non-grid location is formulated as:

$$\boldsymbol{P_{xy}}(x,y) = \sum_{i=0}^{1}\sum_{j=0}^{1} w_{ij}\, \boldsymbol{P_{xy}}(x,y)[x_i, y_j], \tag{2}$$

where $x_0 = \lfloor x \rfloor$, $x_1 = x_0 + 1$, $y_0 = \lfloor y \rfloor$, and $y_1 = y_0 + 1$. Here, $\lfloor \cdot \rfloor$ denotes the floor operator that returns the greatest integer less than or equal to the argument, and w_{ij} represents the bilinear interpolation weights, derived from the fractional offsets of (x,y) relative to its four neighboring grid points (x_0,y_0), (x_0,y_1), (x_1,y_0), (x_1,y_1).

Building on the triplane representation, we develop an MLP decoder $\mathcal{D}$ to reconstruct the future point cloud $\boldsymbol{P_{t+H}} \in \mathbb{R}^{N\times 3}$ by predicting the binary occupancy of sampled voxels. Instead of using the full voxel grid, we sample a balanced query set to improve training efficiency. The query set is constructed from occupied voxels, near-free-space voxels, and far-free-space voxels. Specifically, occupied voxels correspond to raw centers, whereas the others represent regions near or distant from the surface geometry, as determined by a predefined distance threshold. Each query voxel, concatenated with its sinusoidal positional embeddings $\mathbf{PE}(\cdot)$, is passed into the decoder $\mathcal{D}$ to predict the occupancy state. The positional embedding facilitates the representation of high-frequency scene details. Formally, the occupancy prediction of a 3D point $\boldsymbol{p}_i$ is given by:

$$\hat{o}_i = \mathcal{D}(\boldsymbol{f_{p_i}}, \mathbf{PE}(\boldsymbol{p}_i)), \tag{3}$$

where $\hat{o}_i \in [0,1]$ denotes the predicted occupancy probability. Overall, the triplane autoencoder is optimized using the focal loss [32], which mitigates the impact of class imbalance:

$$\mathcal{L}_1 = -\frac{1}{N}\sum_{i=1}^{N} \alpha_{\hat{o}_i}\,(1-\hat{o}_i)^{\gamma}\log(\hat{o}_i), \tag{4}$$

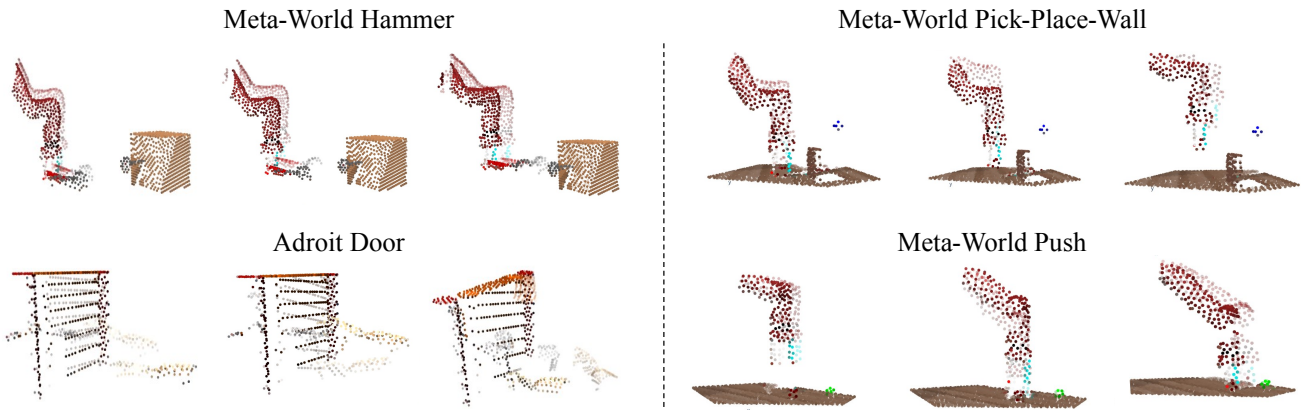

Fig. 3: **Visualization of 3D Imagination on simulation benchmarks**. The raw point clouds (high opacity) and the imagined point clouds (low opacity) are overlaid in the same window to clearly indicate the underlying world dynamics.

where $\alpha_{\hat{o}_i}$ is a class balancing weight, and γ is the focusing parameter. By reducing the influence of abundant occupied and far-free-space samples, this formulation focuses training on the more challenging near-free-space cases.

C. Adaptive Action Generation

The pipeline of the action generation stage is illustrated in Fig. 2 (Bottom). Building upon the learned 3D world model, the objective is to map visual observations O, including the current point cloud $\boldsymbol{P_t} \in \mathbb{R}^{N\times 3}$, the imagined point cloud $\boldsymbol{P_{t+H}} \in \mathbb{R}^{N\times 3}$, and proprioception states $\boldsymbol{S}$, to an action trajectory $\boldsymbol{A_t}$. We consider the imagined 3D point cloud as an additional visual conditional representation. First, point clouds are downsampled using farthest point sampling (FPS) [33] to reduce redundancy. Each modality is encoded by its respective lightweight MLP encoders, after which the resulting visual representations are concatenated and fused into a single conditional feature $\boldsymbol{f_c} \in \mathbb{R}^C$ that jointly guides the subsequent action generation process.

Inspired by [29], [31], we consider a MeanFlow-based method that enables one-step inference and high-quality action generation. Given visual conditions $\mathbf{c}$, the action expert is trained to predict the average velocity $u(\boldsymbol{A}_t, r, t)$, which transports the initially sampled noise $\mathbf{A}_1 \sim \mathcal{N}(0, I)$ into the action trajectory $\boldsymbol{A}_0 = \boldsymbol{A}(t=0)$. Based on the "MeanFlow Identity" and the chain rule for total derivatives, the unknown target velocity u_{tgt} can be expressed in terms of the known instantaneous velocity v_t and the network prediction u_θ:

$$u_{tgt} = v_t - (t-r)\left(v_t \cdot \nabla_{\boldsymbol{A}_t} u_\theta + \partial_t u_\theta\right). \tag{5}$$

To balance the trade-off between sample diversity and conditional consistency, we introduce an adaptive Classifier-Free Guidance (CFG) strategy that linearly combines the conditional and unconditional outputs:

$$v_t^{cfg} \triangleq \omega \cdot v_t\left(\cdot \mid \mathbf{c}\right) + (1-\omega)\cdot u_\theta^{cfg}\left(\boldsymbol{A}_t, t, t \mid \varnothing\right), \tag{6}$$

where the CFG-aware velocity field v_t^{cfg} is obtained by combining the instantaneous velocity $v_t\left(\cdot \mid \mathbf{c}\right)$ under class conditions $\mathbf{c}$ with the unconditional average velocity u_θ^{cfg} predicted by the network. To modulate the contribution of conditional guidance, we introduce a cosine-based gating function with scale ω, defined as:

$$\omega = \begin{cases} 0, & \rho \le \tau_l, \\ \dfrac{\rho - \tau_l}{\tau_h - \tau_l}, & \tau_l < \rho < \tau_h, \\ 1, & \rho \ge \tau_h, \end{cases} \tag{7}$$

where ρ denotes the cosine similarity between v_t and u_θ^{cfg}, and τ_l and τ_h represent predefined lower and upper thresholds. The training objective is finally formulated as the mean squared error (MSE) between the predicted and target velocity fields:

$$\mathcal{L}(\theta) = \mathbb{E}\left\| u_\theta^{cfg}\left(z_t, r, t\right) - \mathrm{sg}\left(u_t\right)\right\|_2^2, \tag{8}$$

where a stop-gradient operation $\mathrm{sg}(\cdot)$ is used to maintain training stability.

IV. Simulation Experiments

A. Benchmark

To evaluate the proposed method, we adopt two widely used and challenging simulators: Adroit [34] and Meta-World [35]. The Adroit benchmark features three complex dexterous manipulation tasks that require fine-grained control of a high-DoF articulated hand. The Meta-World benchmark provides a diverse set of 34 tasks performed with a two-fingered robotic arm, covering a wide range of difficulty levels: 21 Easy, 4 Medium, 4 Hard, and 5 Very Hard tasks. The benchmark facilitates a systematic evaluation of policy precision and stability across diverse task complexities.

B. Baselines

In this work, we conduct a comprehensive comparison of visuomotor policies conditioned on either 2D or 3D representations. The diffusion-based methods include Diffusion Policy (DP) [24], Consistency Policy (CP) [36], and

TABLE I: Performance Evaluation on Adroit and Meta-World Benchmarks.

Methods	Adroit			Meta-World				Average
	Hammer	Door	Pen	Easy (21)	Medium (4)	Hard (4)	Very Hard (5)	
DP	16 ± 10	34 ± 11	13 ± 2	50.7 ± 6.1	11.0 ± 2.5	5.25 ± 2.5	22.0 ± 5.0	35.2 ± 5.3
CP	45 ± 4	31 ± 10	13 ± 6	69.3 ± 4.2	21.2 ± 6.0	17.5 ± 3.9	30.0 ± 4.9	50.1 ± 4.7
DP3	100 ± 0	56 ± 5	46 ± 10	87.3 ± 2.2	44.5 ± 8.7	32.7 ± 7.7	39.4 ± 9.0	68.7 ± 4.7
FlowPolicy	98 ± 1	61 ± 2	54 ± 4	84.8 ± 2.2	58.2 ± 7.9	40.2 ± 4.5	52.2 ± 5.0	71.6 ± 3.5
MP1	100 ± 0	69 ± 2	58 ± 5	88.2 ± 1.1	68.0 ± 3.1	58.1 ± 5.0	67.2 ± 2.7	78.9 ± 2.1
Ours	$\mathbf{100 \pm 0}$	$\mathbf{74 \pm 3}$	$\mathbf{63 \pm 4}$	$\mathbf{93.4 \pm 1.0}$	$\mathbf{74.7 \pm 3.7}$	$\mathbf{65.0 \pm 4.6}$	$\mathbf{74.2 \pm 4.8}$	$\mathbf{84.5 \pm 2.3}$

DP3 [12]. In addition, we evaluate the state-of-the-art flow-based architectures, e.g., FlowPolicy [28] and MP1 [29]. Our world model is highly adaptable and can be integrated into both diffusion- and flow-based architectures.

C. Evaluation Metrics

To evaluate 3D scene prediction, we employ the voxel-based IoU metric, which measures the volumetric overlap between predicted and ground-truth occupancy. For action generation, following [12], the evaluation metric is defined as the average of the Top-5 success rates across 20 randomly sampled episodes for every 200 training epochs. Results are reported as the mean and standard deviation across three independent runs with different random seeds.

D. Implementation Details

All models are deployed on a single NVIDIA RTX 3090 GPU with a batch size of 4 for the triplane autoencoder and 128 for the action expert. For the triplane autoencoder, each input scene is encoded into a triplane representation with spatial resolution (H', W', D') = (128, 128, 32), and the feature dimension C is 16. The threshold for near-free-space is set to 10. For each downstream task, we first collect 10 expert demonstrations from scripted policies [35]. Each observation modality, including the current 3D point cloud, the imagined point cloud, and the robot state, is encoded into a unified 64-dimensional embedding. We use an observation horizon of two steps, where observations from the two most recent frames are incorporated as conditional inputs to the policy. The point cloud is downsampled to 512 points for Adroit and 1024 points for Meta-World using farthest point sampling (FPS). The training process is iterated for 3,000 epochs. Finally, both action and state inputs are normalized to the interval [-1, 1] to stabilize training.

E. Results

In Fig. 3, we visualize the imagined point clouds on the Adroit-Door, Meta-World Hammer, Meta-World Pick-Place-Wall, and Meta-World Push tasks. The raw and imagined results are rendered with different opacities to highlight the underlying world dynamics. A comparison of the tracking motions shows that our model consistently demonstrates the capability to identify dynamic object interactions. For example, in the Meta-World Hammer task, the robot must first move downward to grasp the hammer and then push forward to hit the nail. The generated point clouds accurately capture such motion trajectories, which demonstrate coherent visual semantics and physical consistency. The snapshots of task execution are illustrated in Fig. 4. In the Meta-World Hammer and Assembly tasks, the baseline suffers from severe misalignment and failure, whereas our method achieves stable and precise manipulation during the final contact stage.

Besides qualitative results, we further present quantitative evaluations in Table I. The success rates on the 3 Adroit and 34 Meta-World tasks are reported. Compared with existing diffusion-based (rows 1-3) and flow-based (rows 4-5) methods, our method achieves the best results across all tasks, with an average performance improvement of 5.6% over MP1 and 12.9% over DP3. Notably, it delivers the largest gains on the Meta-World (Very Hard) tasks, where the performance gains become more pronounced as task difficulty increases. These gains arise from the informative guidance of 3D imagined point clouds, which allow the robot to anticipate future states and plan actions more effectively. Incorporating such world models into manipulation policies yields more robust and generalizable performance.

F. Ablation Studies

1) Triplane Representation: We conduct ablation studies to evaluate the triplane representation in the 3D World Imagination stage. We compare it with baselines that use a single xy-plane feature map or full 3D volumetric features. As shown in Table II, the xy-plane baseline achieves moderate IoU performance (e.g., 84.9 on Basketball, 82.2 on Hammer). Using 3D volumes improves accuracy across tasks (up to 92.2 on Hammer), suggesting that richer 3D cues are beneficial. Our triplane representation further boosts performance consistently across all benchmarks (e.g., 95.7 on Push, 94.7 on Hammer). These results highlight that triplanes not only preserve geometric fidelity but also offer a more efficient and effective encoding for 3D scene structure.

2) Adaptive CFG: Table III presents a comparison between the MP1 baseline and our method under settings with and without adaptive CFG. Incorporating this strategy yields consistent gains across all simulation tasks. On the Meta-World benchmark, adaptive CFG improves the average

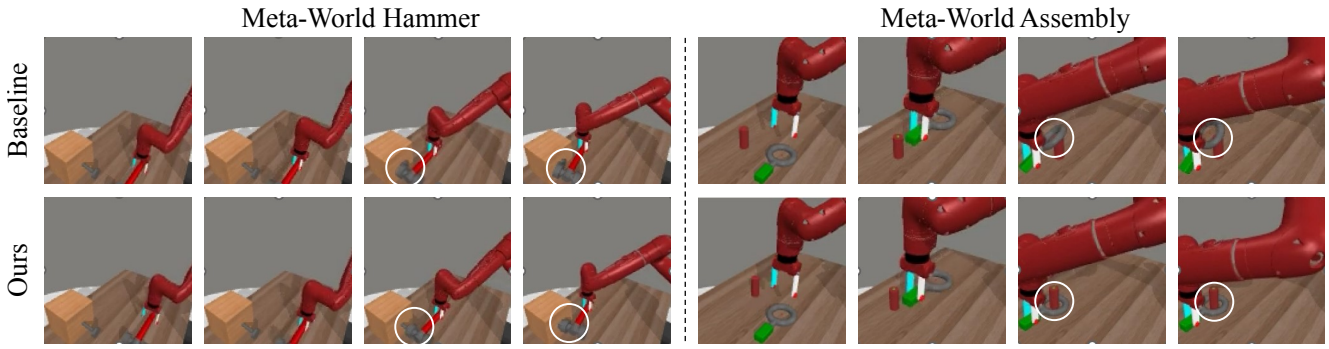

Fig. 4: **Qualitative comparison on the Meta-World Hammer and Assembly tasks.** The baseline suffers from misalignment and task failure (circled), while our method achieves more stable and precise manipulation (circled) across time steps.

TABLE II: Ablation studies on 3D representations.

Method	Basketball	Push	Hammer	Assembly
xy-plane	84.9	86.6	82.2	86.8
3D-volume	89.6	91.1	92.2	90.3
Triplane (Ours)	92.8	95.7	94.7	91.9

TABLE III: Ablation studies on adaptive CFG.

Method	Hammer	Door	Pen	MW-Average
MP1 (w/o A-CFG)	100 ± 0	69 ± 2	58 ± 5	79.2 ± 2.0
MP1 (w/ A-CFG)	100 ± 0	70 ± 2	60 ± 4	80.9 ± 1.8
Ours (w/o A-CFG)	100 ± 0	73 ± 3	60 ± 4	84.1 ± 2.1
Ours (w/ A-CFG)	100 ± 0	74 ± 3	63 ± 4	85.0 ± 2.3

success rate of MP1 from 79.2% to 80.9% (+1.7%), and our method from 84.1% to 85.0% (+0.9%). These gains arise from the balancing effect of adaptive CFG between action generation quality and adherence to visual conditions.

3) Data Efficiency: To examine the amount of training data, we conduct an ablation study on the number of expert demonstrations. As illustrated in Table IV, the success rate consistently increases with the number of demonstrations provided. Even with just one or two demonstrations per task, our method yields substantial improvements, such as increasing success from 1% to 8% and from 7% to 20% in Stick-Pull. When the number of demonstrations exceeds ten, the MP1 baseline begins to saturate, whereas our method attains near-perfect success rates across tasks (e.g., 96% in Push and 88% in Stick-Pull). These results demonstrate that our proposed method scales effectively with larger datasets while remaining robust under data-scarce conditions.

V. Real-world Experiments

A. Task Setup

We conduct real-world experiments with a UR5e robot arm and a parallel-jaw gripper. Multi-view RGB-D observations are provided by a third-person Azure Kinect camera and an eye-in-hand Intel RealSense camera. The evaluation

TABLE IV: Ablation study on data efficiency (MP1 / Ours).

Demons.	Lever-Pull	Assembly	Push	Stick-Pull
1	6/15	8/12	9/17	1/8
2	18/26	11/25	22/35	7/20
5	19/30	80/89	32/50	52/65
10	81/85	98/100	74/85	74/81
20	82/86	100/100	91/96	82/88

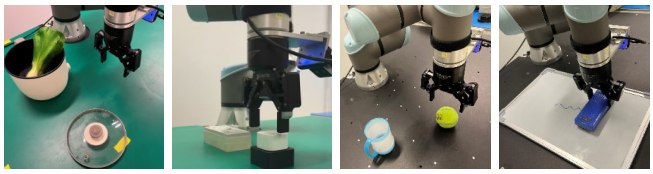

Fig. 5: **Our real-world experimental setup for four tasks: Cooking, Insertion, Pick-Ball, and Wiping.** Multi-view RGB-D observations are provided by an Azure Kinect (third-person) and an Intel RealSense (eye-in-hand).

comprises four tasks: (1) **Cooking**: open the lid and place the vegetable inside, (2) **Insertion**: insert a plug into a socket, (3) **Pick-Ball**: place a tennis ball into a cup, and (4) **Wiping**: clean a designated whiteboard. Fig. 5 illustrates the real-world experimental setup. These tasks span long-horizon planning, contact-rich interactions, and precise object manipulation. For each task, we collect 50 expert demonstrations via teleoperation with a 3Dconnexion SpaceMouse. Based on the simulation results, we select the diffusion-based DP3 and the flow-based MP1 as the real-world baselines. The fused point clouds from both the eye-in-hand and third-person views serve as conditional inputs to the policy.

B. Results

In real-world experiments, each task is evaluated over 15 trials. A trial is considered successful if, in the Cooking task, the robot sequentially opens the lid before placing the vegetable inside, or if, in the Insertion task, it picks up the plug and inserts it accurately into the socket. The quantitative results reported in Table V closely follow the

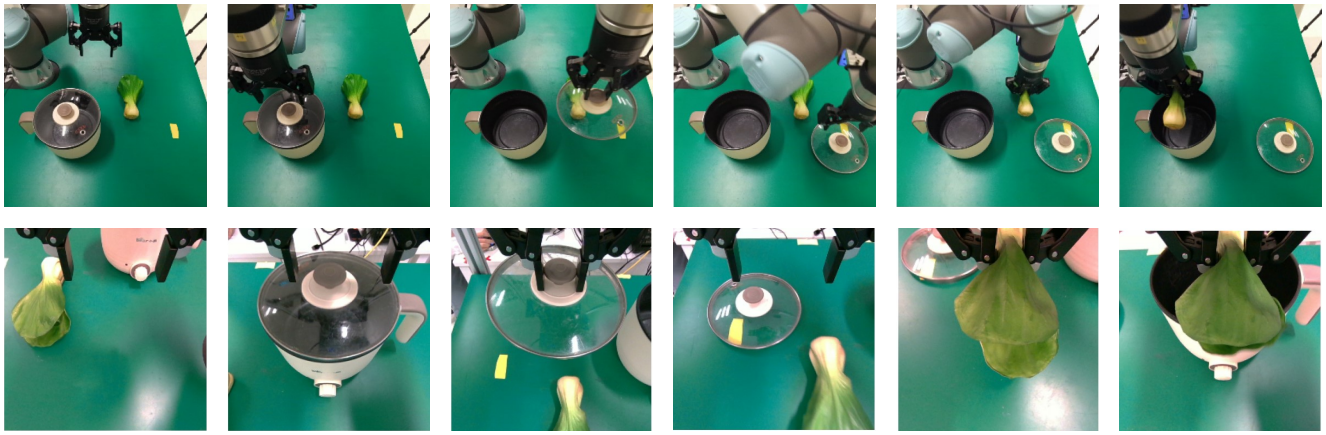

Fig. 6: **RGB snapshots of Cooking.** The third-person **(Top)** and eye-in-hand **(Bottom)** observations during policy execution. The stages are: (1) opening the lid, (2) placing the lid aside, (3) grasping the vegetable, and (4) placing it inside.

TABLE V: Evaluation results on real-world tasks.

Tasks	Cooking	Insertion	Pick-Ball	Wiping
DP3	0/15	3/15	10/15	4/15
MP1	0/15	4/15	11/15	5/15
Ours	4/15	7/15	14/15	8/15

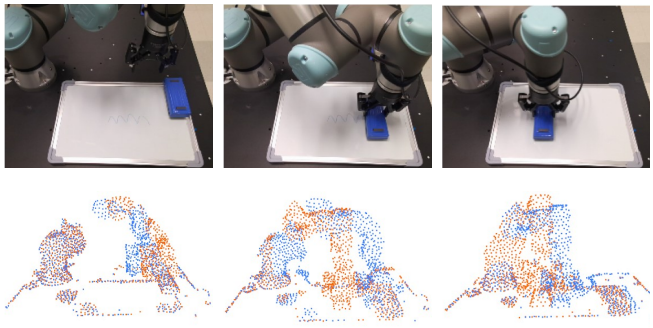

Fig. 7: **Visualization of 3D Imagination in the Wiping task. Row 1:** RGB observations; **Row 2:** Current point clouds (blue) and model-predicted point clouds (orange).

trends observed in simulation experiments. Compared to the DP3 and MP1 baselines, our method consistently achieves higher success rates with only 50 demonstrations. In the Wiping task, our method improves the success rate from $5/15$ to $8/15$, yielding a 20% absolute gain and a 60% relative improvement over the baseline. In the Insertion task, the success rate increases from $4/15$ to $7/15$. Fig. 6 presents the RGB snapshots during our policy execution. The baseline methods fail in all trials, whereas our approach successfully completes the long-horizon lid-opening and placement sequence.

Qualitatively, we visualize the RGB observations together with the predicted future point clouds (yellow) in Fig. 7. Prior to policy execution, our method generates these point cloud predictions and provides a preview of the expected

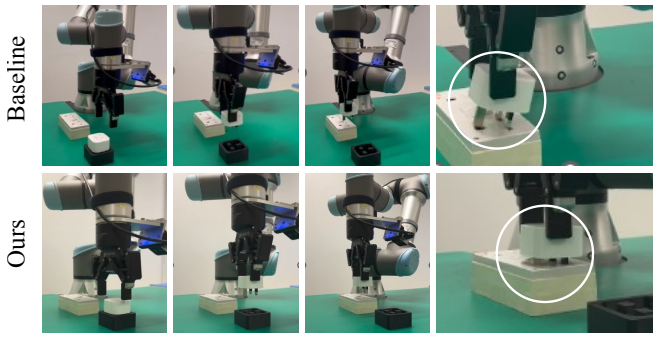

Fig. 8: **Task progress of Insertion.** The baseline **(Top)** suffers from a misalignment failure (circled), whereas our approach **(Bottom)** achieves precise insertion (circled).

scene evolution. In the Wiping task, for instance, the motion trajectory requires the robot arm to first approach downward and then perform a leftward wiping motion. The motion flow between predicted and ground-truth point clouds aligns well with the physical logic of task execution, indicating that our world model faithfully captures task-specific dynamics. Fig. 8 illustrates the progress of Insertion execution. Compared to the MP1 baseline, our policy achieves precise and reliable insertion in the final contact stage. The qualitative and quantitative results demonstrate the stability and effectiveness of our method, highlighting its ability to provide reliable future imagination for complex real-world robotic manipulation.

VI. Conclusion

In this work, we have introduced a hierarchical framework that integrates 3D imagination, perception, and action generation to advance visuomotor policy learning in robotic manipulation. By leveraging a triplane-based 3D world model, our method anticipates future scene dynamics, mitigating the myopic limitation of conventional 2D embodied foundation models. Unlike approaches that only forecast the final outcome state, our world model explicitly predicts

intermediate states of the interaction, providing fine-grained guidance for step-by-step decision-making. The proposed adaptive Classifier-Free Guidance further enhances action generation by balancing fidelity and condition adherence. Extensive experiments on Adroit, Meta-World, and real-world tasks demonstrate the effectiveness, robustness, and generalizability of our approach. Looking ahead, we plan to develop a reward function grounded in the proposed world model to provide feedback for robotic skill acquisition. We believe this work provides a promising step toward scalable, efficient, and cognitively inspired robotic systems capable of reasoning about future states to perform complex manipulation tasks.

REFERENCES

[1] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, "A survey of imitation learning: Algorithms, recent developments, and challenges," *IEEE Transactions on Cybernetics*, 2024.

[2] W. Wang, H. Zhu, and M. H. Ang Jr, "SGSIN: Simultaneous grasp and suction inference network via attention-based affordance learning," *IEEE Transactions on Industrial Electronics*, 2024.

[3] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3153–3160.

[4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[6] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[8] S. Liu, T. J. Teo, Z. Lin, and H. Zhu, "Relationgrasp: Object-oriented prompt learning for simultaneously grasp detection and manipulation relationship in open vocabulary," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 10 890–10 896.

[9] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *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)*. IEEE, 2024, pp. 6892–6903.

[10] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An embodied generalist agent in 3d world," *arXiv preprint arXiv:2311.12871*, 2023.

[11] W. Wang, H. Zhu, and M. H. Ang, "GraspContrast: Self-supervised contrastive learning with false negative elimination for 6-dof grasp detection," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7294–7300.

[12] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D Diffusion Policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[13] D. McNamee and D. M. Wolpert, "Internal models in biological control," *Annual review of control, robotics, and autonomous systems*, vol. 2, no. 1, pp. 339–364, 2019.

[14] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang *et al.*, "WorldVLA: Towards autoregressive action world model," *arXiv preprint arXiv:2506.21539*, 2025.

[15] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3D-VLA: A 3D vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.

[16] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 123–16 133.

[17] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024.

[18] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman *et al.*, "Video generation models as world simulators," *OpenAI Blog*, vol. 1, no. 8, p. 1, 2024.

[19] J. Cho, F. D. Puspitasari, S. Zheng, J. Zheng, L.-H. Lee, T.-H. Kim, C. S. Hong, and C. Zhang, "Sora as an AGI world model? a complete survey on text-to-video generation," *arXiv preprint arXiv:2403.05131*, 2024.

[20] Z. Zhu, X. Wang, W. Zhao, C. Min, N. Deng, M. Dou, Y. Wang, B. Shi, K. Wang, C. Zhang *et al.*, "Is Sora a world simulator? a comprehensive survey on general world models and beyond," *arXiv preprint arXiv:2405.03520*, 2024.

[21] Y. Guan, H. Liao, Z. Li, J. Hu, R. Yuan, G. Zhang, and C. Xu, "World models for autonomous driving: An initial survey," *IEEE Transactions on Intelligent Vehicles*, 2024.

[22] S. Zhou, Y. Du, J. Chen, Y. Li, D.-Y. Yeung, and C. Gan, "RoboDreamer: Learning compositional world models for robot imagination," *arXiv preprint arXiv:2404.12377*, 2024.

[23] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[24] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion Policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[25] M. Zhu, Y. Zhu, J. Li, J. Wen, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng, F. Feng *et al.*, "Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10 838–10 845.

[26] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[27] X. Hu, Q. Liu, X. Liu, and B. Liu, "Adaflow: Imitation learning with variance-adaptive flow-based policies," *Advances in Neural Information Processing Systems*, vol. 37, pp. 138 836–138 858, 2024.

[28] Q. Zhang, Z. Liu, H. Fan, G. Liu, B. Zeng, and S. Liu, "Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 14 754–14 762.

[29] J. Sheng, Z. Wang, P. Li, and M. Liu, "MP1: Mean flow tames policy learning in 1-step for robotic manipulation," *arXiv preprint arXiv:2507.10543*, 2025.

[30] L. Yang, Z. Zhang, Z. Zhang, X. Liu, M. Xu, W. Zhang, C. Meng, S. Ermon, and B. Cui, "Consistency flow matching: Defining straight flows with velocity consistency," *arXiv preprint arXiv:2407.02398*, 2024.

[31] Z. Geng, M. Deng, X. Bai, J. Z. Kolter, and K. He, "Mean flows for one-step generative modeling," *arXiv preprint arXiv:2505.13447*, 2025.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[33] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[34] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv preprint arXiv:1709.10087*, 2017.

[35] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[36] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg, "Consistency Policy: Accelerated visuomotor policies via consistency distillation," *arXiv preprint arXiv:2405.07503*, 2024.