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

3D Dynamics-Aware Manipulation: Endowing Manipulation Policies with 3D Foresight

Yuxin He[1], Ruihao Zhang[1†], Xianzu Wu[1†], Zhiyuan Zhang[1], Cheng Ding[2], Qiang Nie[1⋆]

***Abstract*— The incorporation of world modeling into manipulation policy learning has pushed the boundary of manipulation performance. However, existing efforts simply model the 2D visual dynamics, which is insufficient for robust manipulation when target tasks involve prominent depth-wise movement. To address this, we present a 3D dynamics-aware manipulation framework that seamlessly integrates 3D world modeling and policy learning. Three self-supervised learning tasks (current depth estimation, future RGB-D prediction, 3D flow prediction) are introduced within our framework, which complement each other and endow the policy model with 3D foresight. Extensive experiments on simulation and the real world show that 3D foresight can greatly boost the performance of manipulation policies without sacrificing inference speed. Code is available at https://github.com/Stardust-hyx/3D-Foresight.**

I. INTRODUCTION

An exciting direction for improving language-conditioned manipulation policies is the incorporation of world modeling, i.e., predicting the transition of world states driven by impetus like language command or low-level actions. Recently, this line of work has demonstrated promising results by pretraining policy models on large-scale video data to predict future RGB observation [1]–[5]. Through this kind of 2D world model learning, policy models become aware of the desired future states under current observations and can more easily predict actions that lead to such states.

Unfortunately, the monocular 2D description of the world is lossy in terms of depth information, which is valuable for distance guidance and obstacle avoidance. Yet **it is possible to infer depth well from a monocular image**, as showcased by one-eyed persons or monocular depth estimation neural networks [6]–[8]. This leads to a key insight behind our work: Instead of praying for our models to develop such an ability implicitly, it is more practical to explicitly teach our models about that. Another key insight is that **both the 3D scene transformation and low-level SE(3) robotic actions share the same 3D space and a similar dynamics**, i.e., the underlying trend of how everything should move driven by the same language command. Policy models with 3D foresight will be able to capture this underlying trend and behave accordingly.

Motivated by these two insights, we come up with a 3D dynamics-aware manipulation framework that seamlessly integrates 3D world modeling and policy learning, so as to endow manipulation policies with 3D foresight. In the core of our framework are three complementary self-supervised learning tasks: current depth estimation, future RGB-D prediction, and 3D flow prediction. Cross-embodiment pretraining and downstream fine-tuning are conducted with these auxiliary learning objectives.

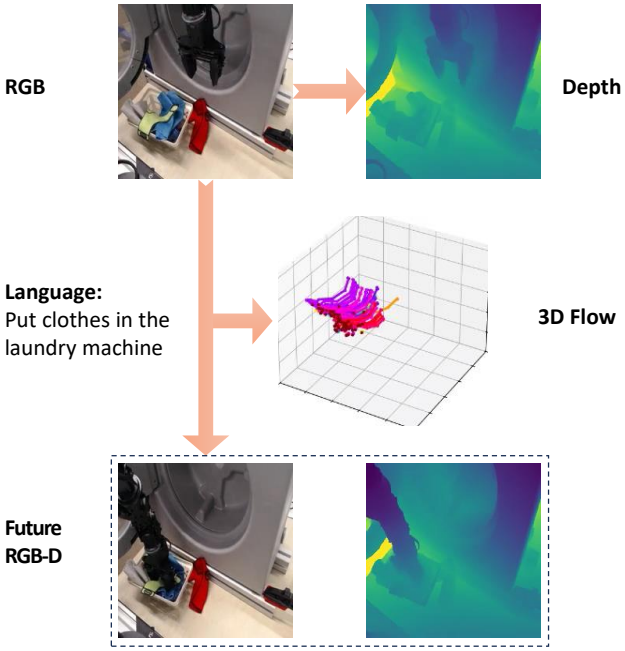

Fig. 1: We propose to equip manipulation policies with 3D foresight by letting the policy models learn to predict depth, future RGB-D and 3D flow from observed RGB. We lift large-scale manipulation demonstration videos into 3D and track the points within to achieve self-supervised pretraining and finetuning.

We choose to represent the 3D world with RGB-D to reduce the burden of data preprocessing. Although it is possible to further lift videos into sequences of point clouds via 3D reconstruction, how to satisfactorily achieve that on large-scale in-the-wild data with limited computing resources is still an open question, which is why we leave it for future research. In order to represent the transformation of 3D scene, we leverage 3D flow [9]–[11] as a bridge between current and future RGB-D frames, as shown in Fig. 1. We find that combining these related factors can let the model learn to calibrate each other during the learning process.

A causal transformer is employed to jointly model the language-driven dynamics of RGB-D, 3D flow and SE(3) robotic actions in an end-to-end manner, maximizing parameter sharing and knowledge transfer. To avoid unnecessary latency, the transformer applies a query-based parallel rep-

[1]The Hong Kong University of Science and Technology (Guangzhou).
[†]Work done during internship. [2]JAKA Robotics Co., Ltd.

resentation updating mechanism. For non-action output, we utilize auxiliary decoding heads to compute corresponding losses during training and remove or offload these auxiliary heads during inference.

Our experiments are carried out on two simulation benchmarks (CALVIN [12], LIBERO [13]) as well as real-world settings, where policy models equipped with 3D foresight achieve state-of-the-art (SoTA) without sacrificing inference speed. In-depth analyses reveal the contribution of 3D foresight to handling tasks that involve prominent depth-wise movement.

In summary, the main contributions of this paper include:

- We propose to endow manipulation policies with 3D foresight by combining 3D world modeling and policy learning under a unified framework.
- Three self-supervised learning tasks (current depth estimation, future RGB-D prediction, 3D flow prediction) are introduced to capture 3D world dynamics.
- The benefit of 3D foresight is verified by experiments on two simulation benchmarks and the real world.

II. RELATED WORK

A. Language-conditioned Manipulation

Instead of constructing a single-task visuomotor policy for every task, it is more desirable to have a multitask language-conditioned manipulation policy that is generalizable. CLIPort [14], which injects frozen language embeddings into two-stream convolution networks, is one of the first studies that show the potential of this paradigm. However, purely relying on frozen language embeddings will impose a limit of generalizability, due to the domain gap between language and vision. Many methods [15]–[18] leverage large vision language models (VLMs) to address this. Despite the promising results that these methods have achieved, they face efficiency and interpretability issues. Another line of work [1]–[5] tries to mine the dynamics hidden in captioned videos, which is known as the world modeling approach [19].

B. World Modeling for Manipulation Policy Learning

Similar to VLM-based methods, the world modeling approach benefits from self-supervised learning on easily accessible unlabeled data. But the world modeling approach is more interpretable, as the core of it is to predict the outcome of behaviors. Existing work [1]–[5] mainly leverages future image generation as the self-supervised learning objective. However, the dynamics of pixels in the 2D frame space is superficial and just loosely related to the dynamics of robotic actions. Our work highlights the significance of modeling the dynamics in 3D space.

C. Flow-enhanced Manipulation Polices

Different from modeling frame-to-frame transition, flow highlights a subset of points within the scene. As a result, movement information is decoupled from visual appearance information, which makes it easier to capture by a model. In [20], 2D flow prediction from the latent space of the policy model acts as an auxiliary learning task, while the predicted flow is not utilized for action guidance. In contrast, all other flow-enhanced manipulation policies [9], [21]–[23] solely employ predicted flow to guide action prediction. Among them, flow is typically confined to the 2D frame space, except for *General Flow* [10] and *G3Flow* [11], which also adopt 3D flow. However, the focus of *General Flow* is developing a 3D flow prediction model and the focus of *G3Flow* is semantic-oriented 3D reconstruction and representation, whereas our work features end-to-end integration of 3D flow.

III. METHOD

A language-conditioned action-chunking policy $\pi_\theta(\mathbf{a}_{t:t+K-1}|\mathbf{o}_{t-T+1:t}, \mathbf{c})$ takes a language command $\mathbf{c}$ and observations $\mathbf{o}_{t-T+1:t}$ (images, proprioception states, etc.) with historical window size T as input, and outputs a chunk of actions $\mathbf{a}_{t:t+K-1}$ of length K. These variables belong to quite distinct modalities, which is why **knowledge from large-scale diverse data and auxiliary learning objectives that narrow the domain gap are necessary**. 3D Foresight incorporates the prediction of current depth $\mathbf{d}_t^{\text{main/wrist}}$, future RGB-D $\mathbf{rgbd}_{t+S}^{\text{main/wrist}}$ in main/wrist views and 3D flow $\boldsymbol{\tau}_{t:t+L-1} \in \mathbb{R}^{L\times P\times 3}$, where S is the time shift of future images, L is the flow length and P is the number of track points. In our formulation, the three dimensions of 3D flow correspond to x, y (in pixel coordinates) and metric depth value, respectively. Our framework employs a causal transformer to model the multimodal spatial-temporal correlation, as shown in Fig. 2.

A. Causal Modeling of Multi-modal Input and Queries

Following previous research [1], [2], the language command is encoded into a vector $\mathbf{c} \in \mathbb{R}^d$ (d is the model dimension) with CLIP [24] text encoder and a linear projection layer. Each main/wrist-view RGB image is encoded into a matrix of $(1 + r)$ vectors via MAE [25] and a perceiver resampler, where the first vector is the *CLS* token and the rest r vectors are resampled from patch tokens. The robot proprioception state, which includes the 6D end effector pose and the binary gripper status, is embedded into a vector $\mathbf{p} \in \mathbb{R}^d$ with linear layers.

The construction of queries for 3D flow, future RGB-D and action chunk is introduced below:

a) Flow Query: The model predicts the future 3D trajectories of grid points (during inference) or randomly sampled points near grids (during training). We initialize l learnable vectors as the flow query. To include information about which points to track, the starting pixel coordinates of sampled points are encoded into a vector via a linear layer, which is added to each flow query vector.

b) Future Query: The query for future RGB-D in main/wrist-view is instantiated as $1 + r$ learnable vectors. The number of future query vectors is equal to the number of image embedding vectors such that appropriate capacity is equipped to reconstruct future RGB-D.

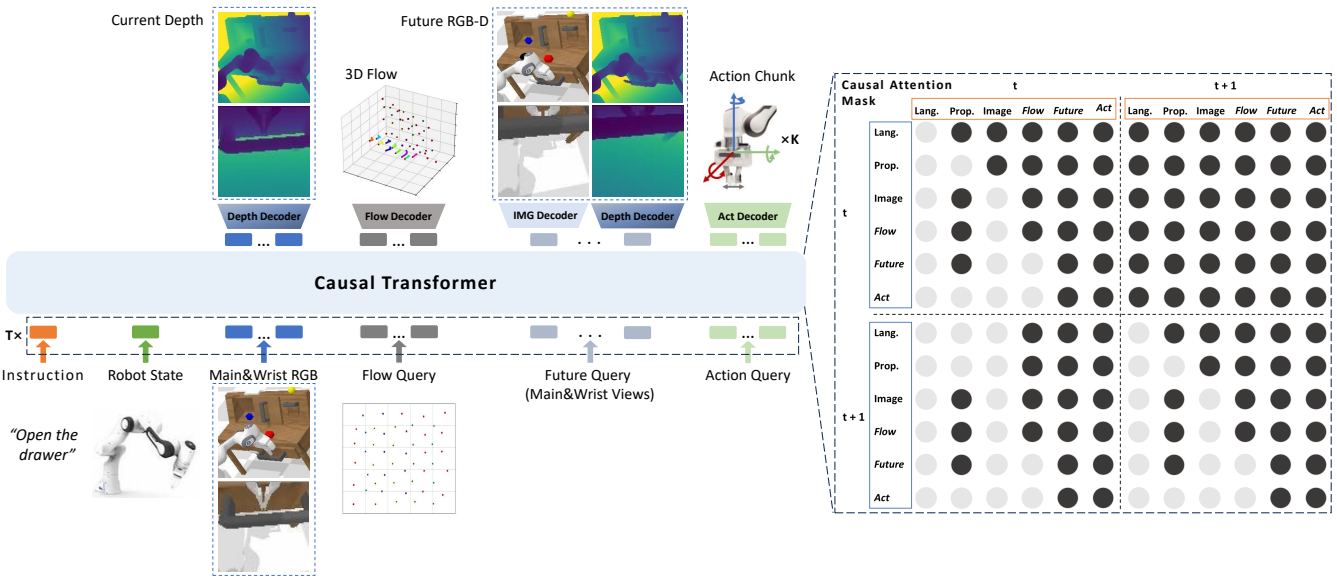

Fig. 2: An overview of the proposed end-to-end framework that captures the language-driven dynamics of 3D world and low-level manipulation actions in a unified manner. In the illustration of the causal attention mask, a light circle mean the horizontal element can attend to the vertical element.

c) Action Query: A learnable vector is responsible for querying action-related information.

The language, vision, proprioception input along with various queries for each timestep are organized into a sequence. And all sequences from timestep $t-T+1$ to timestep t are concatenated and fed into a GPT-style transformer equipped with a carefully designed self-attention mask **in one pass**. Within it, all language, vision, proprioception tokens attend to their historical counterparts to establish temporal relation. All queries attend to current and historical language, vision tokens. Action query additionally attends to current and historical proprioception tokens as well as 3D flow query, since the two are deeply correlated. The hidden states of all tokens are updated by causal self-attention in parallel. Please refer to Fig. 2 for an illustration of the attention mask.

B. 3D World Model Learning and Policy Learning

Three complementary self-supervised learning objectives equip the model with 3D foresight:

1) Current Depth Prediction: We instantiate a depth decoder based on bidirectional self-attention. The input to the depth decoder includes the linear projection of the final hidden states of current main/wrist-view image tokens (acting as context) and a set of 2D masked patch tokens. Here, the masked patch tokens are the sum of 2D sin-cos positional embeddings and a learnable mask vector. The final hidden states of these masked patch tokens are linearly transformed into the predicted current depth.

2) Future RGB-D Prediction: The final representation of future query goes into an image decoder as well as the depth decoder. The image decoder is structurally identical to the depth decoder, except that its output has three channels rather than one channel.

3) 3D Flow Prediction.: The flow decoder works in a similar way as the depth decoder. It first initiates a set of masked patch tokens from starting pixel coordinates, then updates the patch representation conditioned on the final representation of flow query via bidirectional self-attention, then linearly transforms the final patch representation into predicted flow $\hat{\boldsymbol{\tau}}_{t:t+L-1} \in \mathbb{R}^{L\times P\times 3}$.

The model also learns to predict the action chunk in an imitation manner. As in ACT [26] and GR-MG [1], a CVAE encoder compresses the ground-truth action chunk into a latent vector, which are fed into a transformer decoder, along with the final representation of action query and a sequence of positionally-embedded mask tokens. The final hidden states of these mask tokens are transformed into the predicted action chunk $\hat{\mathbf{a}}_{t:t+K-1}$ through an MLP layer. In this paper, we adopt the widely-used SE(3) action space, where an action includes translation (delta of x, y, z), rotation (delta of roll, pitch, yaw) and target binary closeness of the end effector.

The end-to-end learning loss $\mathcal{L}$ over a timestep frame is the combination of prediction losses over current depth, future RGB-D, 3D flow and action chunk:

$$\mathcal{L}_{\text{depth}} = \sum_{v\in\{\text{main,wrist}\}} \text{MSE}(\hat{\mathbf{d}}_t^v, \tilde{\mathbf{d}}_t^v) \tag{1}$$

$$\mathcal{L}_{\text{future}} = \sum_{v\in\{\text{main,wrist}\}} \text{MSE}(\mathbf{rg\hat{b}d}_{t+S}^v, \mathbf{rg\tilde{b}d}_{t+S}^v) \tag{2}$$

$$\mathcal{L}_{\text{flow}} = \text{MSE}(\hat{\boldsymbol{\tau}}_{t:t+L-1}, \boldsymbol{\tau}_{t:t+L-1}) \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\text{act}} &= \text{SmoothL1}(\hat{\mathbf{a}}_{t:t+K-1}[:6], \mathbf{a}_{t:t+K-1}[:6]) \\ &\quad + \alpha\cdot\text{BCE}(\hat{\mathbf{a}}_{t:t+K-1}[6], \mathbf{a}_{t:t+K-1}[6])\end{aligned} \tag{4}$$

$$\mathcal{L} = \beta\cdot\mathcal{L}_{\text{depth}} + \gamma\cdot\mathcal{L}_{\text{future}} + \delta\cdot\mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{act}} \tag{5}$$

where MSE, SmoothL1 and BCE mean Mean Squared

Error, Smooth L1 and Binary Cross Entropy respectively. $\tilde{\mathbf{rgbd}}^v$ and $\tilde{\mathbf{d}}^v$ mean that the pixel-wise normalization [25] is applied to the target value of the R, G, B and depth channels. We set $\alpha = 0.01$ following GR-MG [2], and $\beta = 0.05$, $\gamma = 0.1$, $\delta = 0.1$, which work well consistently throughout three environments (CALVIN, LIBERO, and the real world).

C. Automatic Annotation of Depth and 3D Flow

In many datasets, depth information is not recorded and we use a combination of SoTA monocular metric depth estimator Depth-Anything-V2 [6] and video depth estimator Video-Depth-Anything [8] to obtain temporally consistent metric depth estimate. For any RGB-D video, DELTA [27] can efficiently track the 3D position of points within. We employ a sampling strategy that results in nearly 1250 track points with $\frac{1}{4}$ of them moving in the video and the others staying still. We do so to maintain a trade-off balance between teaching the model the 3D dynamics and narrowing the training-inference gap (since track points are just uniformly sampled from grids during inference).

D. Cross-embodiment Pretraining

We pretrain our model on large-scale cross-embodiment manipulation video data. During pretraining, proprioception states are excluded from the input; actions are excluded from the output; wrist-view depth/images are excluded from the input and output, since these elements are always heterogeneous or missing in different data sources.

E. Implementation Details

We instantiate our models with GR-MG [2] checkpoint pretrained on Ego4D [28] videos. GR-MG is a representative policy model based on 2D world modeling, which shares a similar architecture as ours, except for the depth and 3D flow related parts. Note that, GR-MG additionally consists of a goal image generation module based on image editing, and we keep it for a fair comparison with GR-MG. We set the historical window size T as 10, the future interval S as 3, the action chunk size K as 5 and the length of 3D flow L as 6. Our pretraining is then conducted on 44K trajectories from 5 datasets (RH20T [29], Bridge [30], Berkeley UR5 [31] , Mutex [32] and LIBERO [13]) that cover 4 kinds of embodiments (humans, Franka, UR5, WidowX robots). The pretraining process lasts for 35 epochs, which takes 3 days on 4 NVIDIA 4090 GPUs. After that, we finetune and evaluate our models on downstream manipulation tasks.

IV. EXPERIMENTS

Our experiments aim to answer the following questions:

- **Q1:** Can policy models benefit from 3D foresight?
- **Q2:** Is 3D foresight superior to 2D foresight?
- **Q3:** How do different learning objectives contribute to performance? Are these objectives complementary?
- **Q4:** Does the proposed method work in the real world?

A. Setup

1) Environments: Our simulation experiments are carried out on two benchmarks (CALVIN, LIBERO) and the real world, as shown in Fig. 3. CALVIN [12] encompasses 34 manipulation tasks and 4 different scenes (A, B, C, D). 5K expert trajectories with language instructions are provided for each scene. LIBERO [13] consists of 5 task suites (Spatial, Object, Goal, Long and 90). Each task suite has 10 tasks, except for the last task suite, which has 90 tasks. 50 demonstrations are provided for each task. Our real-world experiments are carried out on a JAKA K-1 7DoF robotic arm with a gripper, a fixed main camera (Orbber Gemini 2L) and a wrist camera (Logitech C922Pro). Two tasks that involve prominent depth-wise movement are considered, including: 1) stack two cups; 2) open a drawer, pick a tape from the drawer and place it on the table then close the drawer. We collect 60 demonstrations for each task using a VR-based teleoperation system.

Ground-truth depth values are available in CALVIN, which are directly used as depth labels during finetuning; in LIBERO and our real-world settings, ground-truth depth values are inaccessible and we obtain the depth labels for training with our preprocessing pipeline.

2) Baselines: We compare our method with the following baselines that *leverage world modeling or 3D information*:

- **3D Diffusion Actor** [33] enhances Diffusion Policy with 3D scene representations.
- **RoboUniView** [34] enhances RoboFlamingo with view-invariant 3D representations based on 3D occupancy.
- **ATM** [9] incorporates 2D flow predicted by a track transformer to guide a transformer policy.
- **GR-1** [1] is a GPT-style policy pretrained with the future RGB prediction task on large-scale video data.
- **SeeR** [3] enhances GR-1 by increasing the model size and combining video pretraining with inverse dynamics.
- **GR-MG** [2] enhances GR-1 with action chunking and goal images generated by an image editing model.
- **UP-VLA** [4] a VLA model that combines multimodal understanding and future RGB prediction.
- **2D Foresight (scratch)** is a 2D counterpart of our method that enhances GR-MG by combining future RGB prediction with 2D flow (without cross-embodiment pretraining).
- **3D Foresight (scratch)** enhances GR-MG with the proposed 3D world modeling objectives (without cross-embodiment pretraining).

We do not compare with UniVLA [5] and DreamVLA [35] because they are based on much (at least 5×) larger backbones and it is unfair to compare with them.

B. Can Policy Models Benefit from 3D Foresight? (Q1)

Main results on CALVIN are shown in Table I. Without pretraining, 3D foresight boosts the performance of GR-MG from 3.84 to 4.01 in the in-domain setting (D→D) and from 4.04 to 4.15 in the zero-shot scene transfer setting (ABC→D). This suggests that 3D foresight is directly

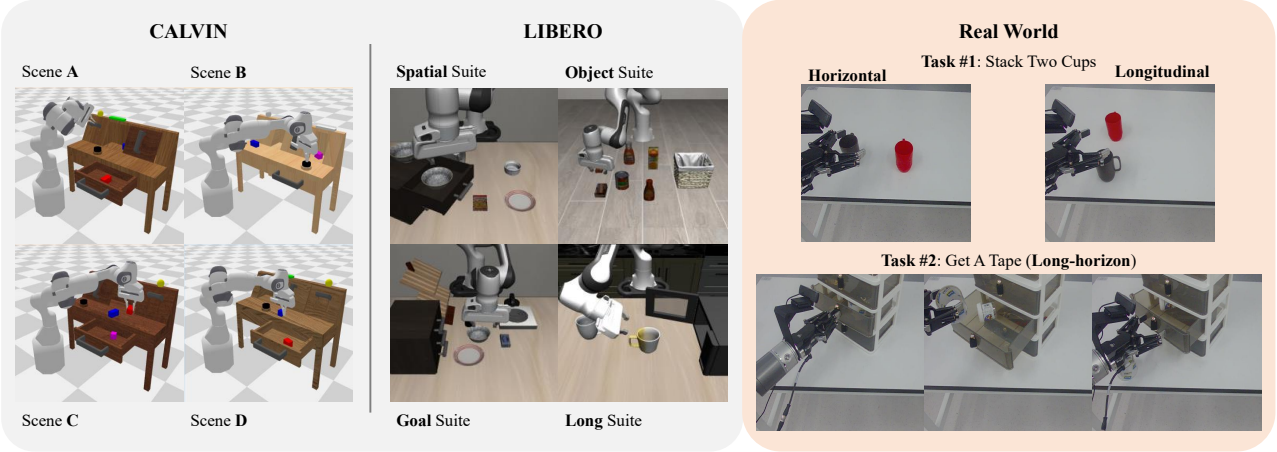

Fig. 3: Environments for our experiments. CALVIN involves 4 scenes of different colors, textures and object placements and 34 manipulation tasks. LIBERO features 4 evaluation task suites that challenge different dimensions of capability. Our real-world setups involve two tasks that require strong spatial awareness.

TABLE I: Overall performance comparison on CALVIN. During evaluation, 1000 chains of tasks are randomly sampled, and the success rates of consecutive 5 tasks are recorded and averaged over all chains. "1" $\sim$ "5" indicate the average success rates of completing 1 $\sim$ 5 tasks. "Avg. Len." means the average number of completed tasks.

Method	D $\rightarrow$ D						ABC $\rightarrow$ D					
	1	2	3	4	5	Avg. Len.	1	2	3	4	5	Avg. Len.
3D Diffusion Actor [33]	-	-	-	-	-	-	93.8	80.3	66.2	53.3	41.2	3.35
RoboUniView [34]	96.2	88.8	77.6	66.6	56.3	3.85	94.2	84.2	73.4	62.2	50.7	3.64
GR-1 [1]	-	-	-	-	-	-	85.4	71.2	59.6	49.7	40.1	3.06
SeeR (base) [3]	-	-	-	-	-	-	94.4	87.2	79.9	72.2	64.3	3.98
UP-VLA [4]	-	-	-	-	-	-	92.8	86.5	81.5	76.9	69.9	4.08
GR-MG [2]	93.0	84.5	76.5	69.0	60.8	3.84	96.8	89.3	81.5	72.7	64.4	4.04
2D Foresight (scratch)	94.8	85.1	76.6	70.8	62.5	3.90	95.6	90.5	83.1	74.8	64.2	4.08
3D Foresight (scratch)	95.5	87.6	80.9	73.2	64.1	4.01	96.2	91.1	84.4	77.1	68.7	4.15
3D Foresight	95.7	88.0	81.6	74.9	66.3	4.08	96.9	92.0	85.7	78.8	71.3	4.23

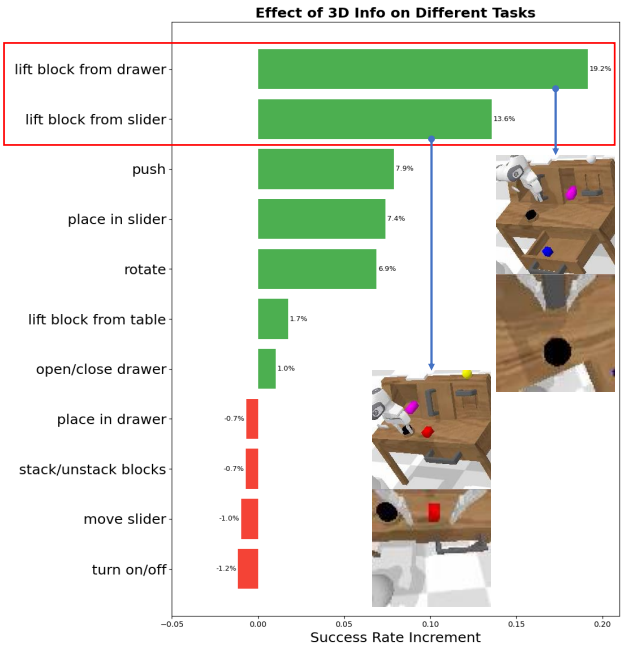

Fig. 4: The increment of task-wise succuss rates on CALVIN D$\rightarrow$D after evolving from 2D Foresight to 3D Foresight.

TABLE II: Overall Performance Comparison on LIBERO

Method	Spatial	Object	Goal	Long	Avg. SR
ATM [9]	84.0	89.4	79.6	65.2	79.6
GR-1 [1]	93.4	93.4	89.0	84.2	90.1
GR-MG [2]	94.0	94.8	91.6	86.4	91.7
2D Foresight (scratch)	94.4	95.8	93.0	87.2	92.6
3D Foresight (scratch)	95.8	97.4	94.0	90.2	94.3
3D Foresight	96.4	98.0	94.8	92.0	95.3

beneficial for manipulation and the benefit will be greater if the target scene is seen during training. With pretraining, the model performance further increases by 0.07 in the in-domain setting and by 0.08 in the zero-shot scene transfer setting, demonstrating the advantage of our framework in leveraging cross-embodiment manipulation videos.

Table II displays the evaluation results on LIBERO. In contrast to CALVIN, LIBERO does not provide ground-truth depth values, and the depth labels for training are obtained with our preprocessing pipeline. Despite that, 3D foresight consistently increases the performance of GR-MG by large margins across the 4 diverse task suites with or without pretraining. This indicates the robustness of our

TABLE III: Ablation Study on CALVIN D→D and the Real World

Method	CALVIN D → D				Real-World		
	$\mathcal{L}_{\text{depth}}$	$\mathcal{L}_{\text{future}}$	$\mathcal{L}_{\text{flow}}$	**Avg. Len.**	$\mathcal{L}_{\text{depth}}$	$\mathcal{L}_{\text{future}}$	$\mathcal{L}_{\text{flow}}$
3D Foresight (scratch)	0.021	0.11	6.8e-5	4.01	0.043	0.20	3.9e-4
w/o Current Depth	-	0.12	7.5e-5	3.97	-	0.23	4.4e-4
w/o Future RGB-D	0.027	-	8.4e-5	3.94	0.048	-	4.8e-4
w/o 3D Flow	0.024	0.14	-	3.95	0.047	0.25	-

TABLE IV: Inference Speed Comparison

Method	Inference Latency
RoboUniView [34]	105 ms
ATM [9]	38 ms
GR-1 [1]	35 ms
UP-VLA [4]	252 ms
GR-MG [2]	106 ms
3D Foresight	112 ms (+6 ms)

TABLE V: Performance Comparison on the Real World

Method	Stack Two Cups		Get A Tape
	Horizontal	**Longitudinal**	
ATM	75%	50%	40%
GR-MG	70%	40%	55%
2D Foresight (scratch)	75%	35%	60%
3D Foresight (scratch)	75%	60%	65%
3D Foresight	80%	70%	75%

method to noisy pseudo depth labels and the versatility of our method to different scenes that challenge different aspects of manipulation ability.

It is worth mentioning that 3D foresight provides all the benefits with only a negligible increase in inference cost. We compare the inference latency of the baseline methods and our method in Table IV. It turns out that our model is only 6 ms slower than the backbone GR-MG. This is achieved by removing or offloading the auxiliary decoding heads for current depth, future RGB-D and 3D flow during inference, since these prediction output are only the by-product of our self-supervise learning objects.

C. 2D Foresight vs. 3D Foresight (Q2)

To compare the effect of 2D foresight and the effect of 3D foresight in a more rigorous way, we implement a strict 2D counterpart of our model. Concretely, we equip GR-MG with a 2D flow prediction module, where the depth dimension is ablated. According to Table I and Table II, with the 2D flow prediction module, the performance of GR-MG increases from 3.84 to 3.90 (+0.06) in CALVIN D→D, from 4.04 to 4.08 in CALVIN ABC→D (+0.04), and from 91.7 to 92.6 (+0.09) in LIBERO, but still lags far behind the proposed 3D foresight method by 0.11, 0.07 and 0.17. The results demonstrate that the advantage of our method does not simply root in the integration of flow prediction, but the comprehensive integration of 3D world modeling.

A more in-depth analysis of the superiority of 3D foresight over 2D foresight is carried out on CALVIN D→D, where we calculate the success rate increments of different manipulation tasks. As shown in Fig. 4, two outstanding tasks that benefit the most from 3D foresight are "lift block from drawer" and "lift block from slider", both of which involve prominent depth-wise movement.

D. Ablation Study (Q3)

To investigate the contributions of different self-supervised learning objectives and to verify the complementarity of these objectives, we conduct a fine-grained ablation study and the results are presented in Table III. It is observed that, in CALVIN D→D, removing future RGB-D prediction and 3D flow prediction reduces the average number of solved tasks by 0.07 and 0.06, respectively, while removing current depth prediction only leads to a decrease by 0.04, indicating that the contribution of dynamics-related objectives is more profound. In addition, removing any of the three learning objectives increases the losses of the other two objectives in the validation set of CALVIN D→D as well as the validation set of our real-world data. This confirms that the proposed self-supervised learning objectives can benefit from each other.

E. Real-world Evaluation and Case Study (Q4)

We carry out real-world experiments to verify the effectiveness of our method in more complex and noisy scenarios. Two tasks (stack two cups, get a tape from the middle drawer) are considered, both of which require strong spatial awareness. For the first task (stack two cups), we design a special evaluation configuration, which involves two placement settings: 1) **Horizontal**, the cups are placed horizontally; 2) **Longitudinal**, the cups are placed longitudinally (see Fig. 3). According to the evaluation results shown in Table V, 3D foresight consistently improve policy performance throughout the tasks. And the performance gain is most prominent in the "stack two cups" task under the longitudinal setting, which involves the most noticeable depth-wise movement. This phenomenon aligns well with our hypothesis.

To qualitatively analyze the performance gain brought by 3D foresight, we take a deeper look into two cases where the 3D Foresight policy succeeded but the 2D Foresight policy failed. The rollouts along with the auxiliary predictions are visualized in Fig. 5. In the first case, the 2D Foresight policy failed to accurately locate the position of the target cup (the wrist view was occluded and the model had to rely on distance perception over the main view) and released the held cup about 6 cm in front of the target cup. In contrast, the 3D Foresight policy successfully located the target cup, thanks to its stronger ability to perceive depth. In the second case, the policies had to determine the timing to grip the drawer handle based on the distance from the gripper to the handle. Different from the first case, the wrist view played a major role in the second case. And it turns out that the 3D Foresight policy does better at perceiving depth from the wrist view, as we explicitly teach that during training.

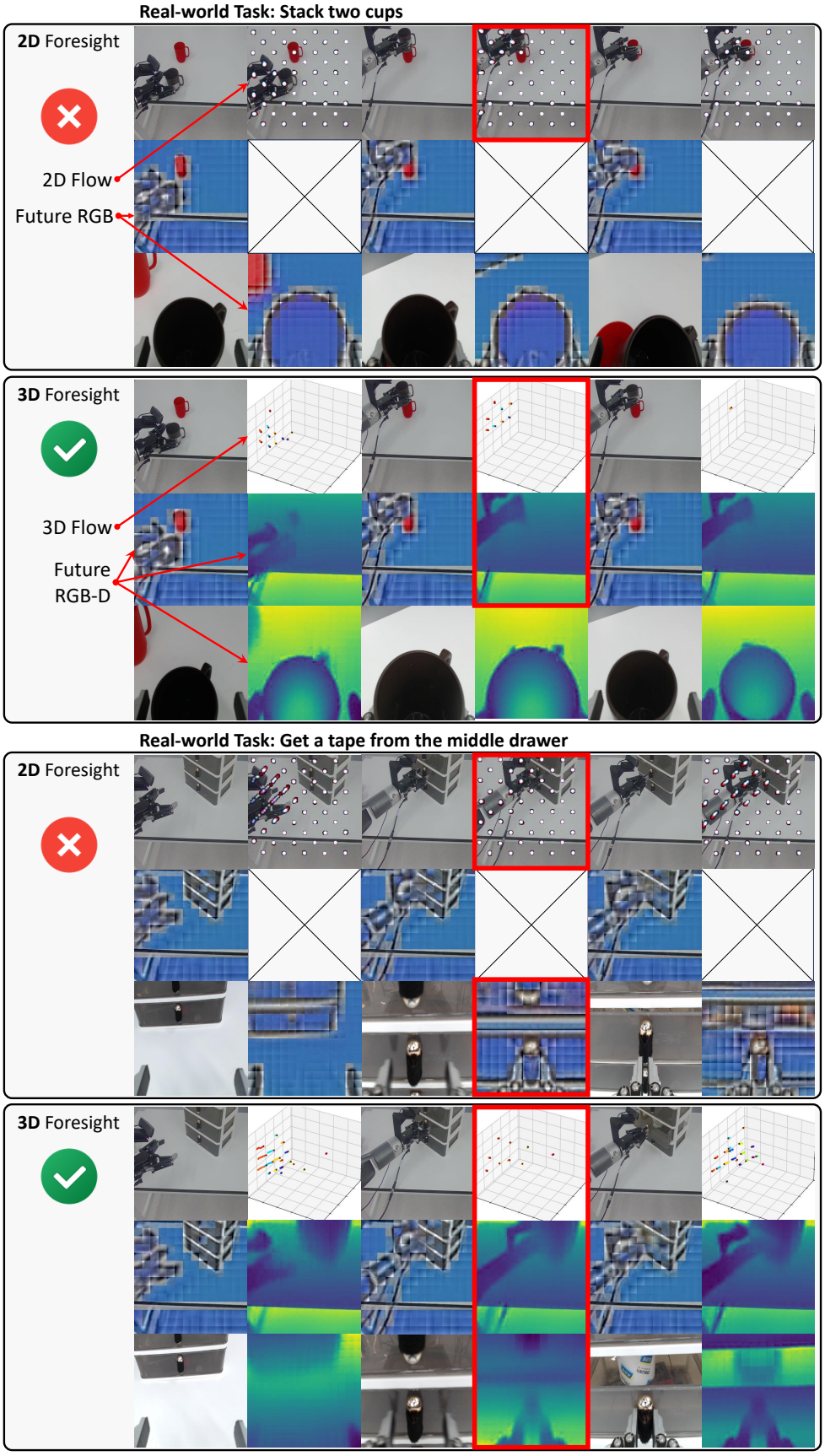

Fig. 5: Visualization of two cases where the 3D Foresight policy succeeded but the 2D Foresight policy failed. Pixel-wise normalization is applied over future RGB during training, which is why the pictures here do not look like normal RGB.

V. CONCLUSION

We explore a way towards 3D dynamics-aware robotic manipulation in this paper. To enhance manipulation policies with 3D foresight, a novel framework is introduced, which integrates 3D world modeling and policy learning through three auxiliary objectives — current depth estimation, future RGB-D prediction, and 3D flow prediction. Experiments across simulation and the real world demonstrate that policies can benefit much more from 3D foresight than from 2D foresight. Our ablation study reveals the complementarity of the proposed learning objectives and the major contribution of 3D dynamics-related objectives to performance. Our case study suggests that 3D foresight can provide manipulation policies with stronger spatial awareness which are critical for tasks that require distance perception or involve prominent depth-wise movement. A promising direction for future work is to explore more advanced 3D scene representations to further enhance the model's spatial reasoning capabilities.

VI. ACKNOWLEDGMENT

This work is in part supported by the Guangzhou-HKUST(GZ) Joint Funding Program (2025A03J3656) , in part supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515012291).

REFERENCES

[1] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," in *ICLR*, 2024.

[2] P. Li, H. Wu, Y. Huang, C. Cheang, L. Wang, and T. Kong, "Gr-mg: Leveraging partially annotated data via multi-modal goal conditioned policy," *arXiv preprint arXiv:2408.14368*, 2024.

[3] Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang, "Predictive inverse dynamics models are scalable learners for robotic manipulation," *arXiv preprint arXiv:2412.15109*, 2024.

[4] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "Up-vla: A unified understanding and prediction model for embodied agent," *ArXiv*, vol. abs/2501.18867, 2025.

[5] Y. Wang, X. Li, W. Wang, J. Zhang, Y. Li, Y. Chen, X. Wang, and Z. Zhang, "Unified vision-language-action model," *arXiv preprint arXiv:2506.19850*, 2025.

[6] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.

[7] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv preprint arXiv:2410.02073*, 2024.

[8] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang, "Video depth anything: Consistent depth estimation for super-long videos," *arXiv preprint arXiv:2501.12375*, 2025.

[9] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," *arXiv preprint arXiv:2401.00025*, 2023.

[10] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," *arXiv preprint arXiv:2401.11439*, 2024.

[11] T. Chen, Y. Mu, Z. Liang, Z. Chen, S. Peng, Q. Chen, M. Xu, R. Hu, H. Zhang, X. Li *et al.*, "G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation," *arXiv preprint arXiv:2411.18369*, 2024.

[12] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 3, pp. 7327–7334, 2022.

[13] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*. PMLR, 2022, pp. 894–906.

[15] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[16] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[17] 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.

[18] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π0: A vision-language-action flow model for general robot control, 2024," *URL https://arxiv.org/abs/2410.24164*, 2024.

[19] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," *Advances in neural information processing systems*, vol. 31, 2018.

[20] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani, "Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation," *arXiv preprint arXiv:2409.16283*, 2024.

[21] M. Vecerik, C. Doersch, Y. Yang, T. Davchev, Y. Aytar, G. Zhou, R. Hadsell, L. Agapito, and J. Scholz, "Robotap: Tracking arbitrary points for few-shot visual imitation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[22] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation," *arXiv preprint arXiv:2405.01527*, 2024.

[23] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," in *8th Annual Conference on Robot Learning*, 2024.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[25] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[26] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[27] T. D. Ngo, P. Zhuang, C. Gan, E. Kalogerakis, S. Tulyakov, H.-Y. Lee, and C. Wang, "Delta: Dense efficient long-range 3d tracking for any video," *arXiv preprint arXiv:2410.24211*, 2024.

[28] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022.

[29] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, "Rh20t: A robotic dataset for learning diverse skills in one-shot," in *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[30] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[31] L. Y. Chen, S. Adebola, and K. Goldberg, "Berkeley UR5 demonstration dataset," https://sites.google.com/view/berkeley-ur5/home, 2023.

[32] R. Shah, R. Martín-Martín, and Y. Zhu, "Mutex: Learning unified policies from multimodal task specifications," *arXiv preprint arXiv:2309.14320*, 2023.

[33] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[34] F. Liu, F. Yan, L. Zheng, C. Feng, Y. Huang, and L. Ma, "Robouniview: Visual-language model with unified view representation for robotic manipulation," *arXiv preprint arXiv:2406.18977*, 2024.

[35] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, H. Wang, Z. Zhang, L. Yi, W. Zeng, and X. Jin, "Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge," *arXiv preprint arXiv:2507.04447*, 2025.