Reimagination with Test-time Observation Interventions: Distractor-Robust World Model Predictions for Visual Model Predictive Control

Yuxin Chen*,1, Jianglan Wei*,1, Chenfeng Xu[1], Boyi Li[1], Masayoshi Tomizuka[1], Andrea Bajcsy[2], Ran Tian[1]
*Denotes equal contribution
[1]*University of California, Berkeley* [2]*Carnegie Mellon University*

***Abstract*—World models enable robots to "imagine" future observations given current observations and planned actions, and have been increasingly adopted as generalized dynamics models to facilitate robot learning. Despite their promise, these models remain brittle when encountering novel visual distractors such as objects and background elements rarely seen during training. Specifically, novel distractors can corrupt action outcome predictions, causing downstream failures when robots rely on the world model imaginations for planning or action verification. In this work, we propose Reimagination with Observation Intervention (ReOI), a simple yet effective test-time strategy that enables world models to predict more reliable action outcomes in open-world scenarios where novel and unanticipated visual distractors are inevitable. Given the current robot observation, ReOI first detects visual distractors by identifying which elements of the scene degrade in physically implausible ways during world model prediction. Then, it modifies the current observation to remove these distractors and bring the observation closer to the training distribution. Finally, ReOI "reimagines" future outcomes with the modified observation and reintroduces the distractors post-hoc to preserve visual consistency for downstream planning and verification. We validate our approach on a suite of robotic manipulation tasks in the context of action verification, where the verifier needs to select desired action plans based on predictions from a world model. Our results show that ReOI is robust to both in-distribution and out-of-distribution visual distractors. Notably, it improves task success rates by up to 3× in the presence of novel distractors, significantly outperforming action verification that relies on world model predictions without imagination interventions.**

I. Introduction

World models [7, 27] enable robots to predict action-conditioned future evolutions of their environments given current observations and planned actions. They have emerged as powerful generalized dynamics models and are increasingly used in model-based policy learning [14, 4, 21] as well as in deployment-time action plan verification [6, 22]. However, despite their promising potential, current world models in robotics are brittle against visual distractors such as task-irrelevant objects or background elements rarely seen during training. These novel distractors can corrupt the model, leading to hallucinated imaginations that ultimately compromise downstream action plan verification and selection. Notably, this brittleness persists even when models are trained on large, visually diverse datasets sourced beyond robotics [1, 26].

To illustrate this challenge, consider a robot assistant tasked with meal preparation, picking up prepped ingredients and placing them into a cooking pan. The robot uses a world model, trained on demonstrations from similarly structured kitchens, to verify candidate action plans proposed by a pre-trained imitation policy. Specifically, the robot uses the world model to imagine the future outcomes of each action plan and evaluates them using a visual reward function to identify the best action plan (i.e., policy verification). However, in open-world deployment, novel visual distractors are inevitable. For example, a user might place an unfamiliar high-pressure pot between the cutting board and the pan, or leave an Amazon package box in the background as illustrated in Figure 1. As the world model rolls out candidate action plans, these unfamiliar distractors often degrade in visually implausible ways, becoming distorted, disappearing, or warping unnaturally across predicted frames. These artifacts cause the model to misrepresent distractors that are critical for ensuring safety and hallucinate incorrect robot behaviors. As a result, the downstream plan verifier failed to detect that a planned motion that would collide with the pot, ultimately leading to deployment-time failures (Figure 1, top).

These failure cases expose a limitation in current world model–based robot policy verification pipelines and motivate the need for strategies that can mitigate the effects of novel distractors. Our key insight is that relying solely on training-time solutions is insufficient: the diversity and unpredictability of distractors, especially in open-world settings, make it impractical to fully capture them, even with large-scale datasets. To this end, we propose Reimagination with Observation Intervention, **ReOI**, a *test-time* and *plug-in* strategy that mitigates the impact of novel visual distractors on world model-based robot policy verification. Our key idea is to first identify and inpaint novel distractors from the current observation to bring it closer to the world model's training distribution, then reimagine future action outcomes using the modified input, and finally reintroduce the distractors post-hoc at the pixel level to preserve visual consistency for downstream planning and verification.

We validate our approach on robotic manipulation tasks in the context of action plan verification, where a verifier needs to select action plans based on visual outcome predictions from a

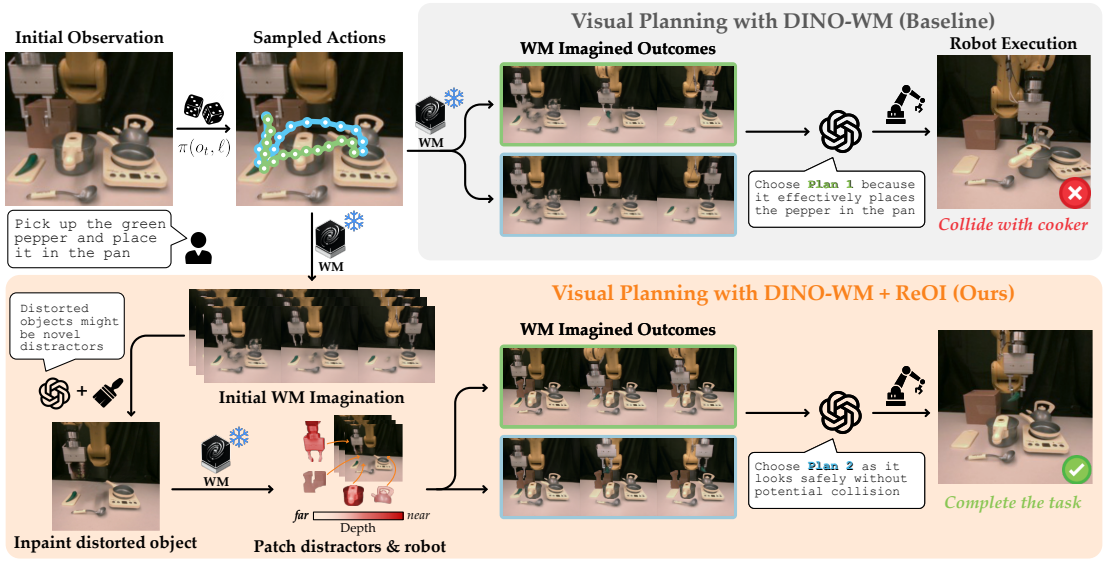

Fig. 1: **Reimagination with observation intervention for distractor-robust world model-based robot planning**. A robot uses a world model (WM) to verify and select action plans proposed by a pre-trained imitation policy. The world model has never encountered the box, cooker, or teapot during training. **Top (baseline)**: Novel visual distractors become distorted across predicted observations, causing the model to hallucinate incorrect robot behaviors and erase a critical obstacle. This leads the verifier to select an unsafe action plan. **Bottom (ours)**: Test-time observation intervention enables the world model to generate predictions better aligned with reality, allowing the verifier to recognize potential collisions and correctly select a safe action plan.

world model. Our approach demonstrates strong robustness to both in-distribution and unfamiliar visual distractors. Notably, our method improves task success rates by up to **3x** when encountering novel distractors compared to standard policy verification that directly relies on unmodified world model predictions.

Contribution. In this work, we raise and tackle the challenge of mitigating the effects of novel visual distractors on world model-based robot policy verification. We illustrate concrete failure cases induced by such distractors and introduce Reimagination with Observation Intervention (**ReOI**), a *test-time*, *plug-in* strategy that enables world models to produce more reliable action outcome predictions in open-world scenarios where novel and unanticipated distractors are inevitable. To the best of our knowledge, this is the first work that leverages test-time observation intervention to address the problem of novel visual distractors in world model–based robot policy verification and selection.

II. Related Works

Mitigating visual distractors in world model learning. Learning accurate world models in visually cluttered environments remains a challenge in robotics. Most existing approaches mitigate visual distractors through training-time strategies: either by leveraging privileged reward signals to learn task-relevant representations [24, 5, 18, 28], or by training separate network branches to model task-relevant and task-irrelevant components, using only the task-relevant branch during downstream policy planning or verification [17, 19, 10]. However, these methods primarily target in-distribution visual distractors and are tightly coupled to the training-time task context. When the task context shifts at test time, causing previously irrelevant distractors to become safety-critical or introducing novel distractors, their performance often degrades and even fails entirely [9]. *Rather than relying on task-specific supervision to suppress distractors during world model training and assuming the same context will hold at deployment, we propose a test-time observation intervention strategy to mitigate novel visual distractors' impact on world model–based robot policy verification.*

Observation intervention for improving robot policy robustness. Modifying robot observations by masking out the background or task-irrelevant objects has proven effective in improving the robustness of *imitation policies* to visual distractors. Prior works have explored using separately trained models [12, 16], VLMs [23], or conformal prediction techniques [8] to identify and inpaint visual distractors to improve policy performance. *Different from these works that focus on modifying observation to improve model-free imitation policy robustness, our work focuses on test-time observation intervention for improving the reliability of model-based visual planning in the presence of novel visual distractors.*

III. Preliminaries

Sampling-based visual model predictive control. We formulate the robot planning problem as a sampling-based visual model predictive control problem where the robot uses

a predictive world model to forward-simulate multiple action-conditioned futures and evaluates them using a reward function to select the best plan for execution. Mathematically, this problem is formulated as:

$$\mathbf{a}_t^\star = \text{argmax}_{\mathbf{a}_t \sim \pi(o_t, \ell)} \mathbb{E}_{\mathbf{o}_t \sim f_\phi(o_t, \mathbf{a}_t)} \Big[R\big(\mathbf{o}_t; \ell\big) \Big], \tag{1}$$

where $f_\phi(o_t, \mathbf{a}_t)$ is a visual dynamics model that predicts a sequence of future observations ($\mathbf{o}_t := o_{t:t+T}$) conditioned on an action plan $\mathbf{a}_t$ and the current observation o_t, and R is a reward function that evaluates the predicted outcomes given the task description (ℓ). This framework is particularly appealing because it can leverage pre-trained generative imitation-based policies (π) as action plan samplers and align the robot's behavior with deployment-time task context and preferences without fine-tuning or modifying the base policy.

World model as visual dynamics for action plan verification and selection. World models enable robots to "imagine" future observations given current observations and planned actions. Typically, a world model consists of three key components: an observation encoder model $z_t = \mathcal{E}_\phi(o_t)$ that maps a visual observation o_t into a latent state z_t, a forward dynamics model $z_{t+1} = \tilde{f}_\phi(z_t, a_t)$ that predicts the next latent state conditioned on the current latent state and action, and an observation decoder model $o_t = \mathcal{Q}_\phi(z_t)$ that decodes a latent state to visual observation. World models have been increasingly adopted as generalized dynamics models to produce action-conditioned future outcomes for downstream robot policy learning [21, 27] or verification under the model predictive control framework described in (1) [6, 14, 22].

Challenge: novel visual distractors cause hallucinations and compromise planning. While world models enable robots to imagine future visual outcomes and inform policy learning, the accuracy and confidence about their imaginations heavily depend on the consistency between the training environment and the deployment environment. The existing effort on exploring world models for robotics applications mostly focuses on training-testing consistent conditions [6, 14, 22]; however, real-world deployment inevitably involves open-world environments where novel visual distractors (such as objects and background elements rarely seen during training) are inevitable. Current world models are brittle under such conditions: these novel distractors can corrupt the model, leading to hallucinated imaginations that ultimately compromise downstream action plan verification and selection (refer to the motivating example in Figure 1).

IV. World Model Reimagination with Test-time Observation Intervention for Visual Model Predictive Control

We propose Reimagination with Observation Intervention (**ReOI**), a ***test-time*** and ***plug-in*** strategy that mitigates the impact of novel visual distractors on world model-based policy verification and selection, where a verifier needs to select desired action plans based on visual outcome predictions from a world model.

Since mitigating dynamic distractors remains a largely open and underexplored challenge, in this work, we focus on mitigating the impact of *static* novel visual distractors on world model–based robot planning. We define visual distractors as static objects and background elements that are not directly related to the task from the task specification. Different from the definition in [8], we do not assume that distractors do not affect the robot's motion to reach the goal. Our key idea is to first identify and inpaint problematic novel distractors from the current observation to bring it closer to the training distribution, then reimagine future outcomes using the modified input, and finally reintroduce the distractors post-hoc to preserve visual consistency for downstream planning and verification.

A. Observation Intervention Strategy

Identify novel visual distractors from world model predictions. Since the original training data is typically unavailable at deployment time, we identify novel visual distractors using only the world model's predicted observations. Our key insight is that visual distractors underrepresented in the training distribution tend to exhibit physically implausible behavior in world model rollouts: although initially visible, they quickly become distorted, disappear, or warp unnaturally as the rollout progresses, even when the input actions are in the distribution (shown in Figure 2). This phenomenon arises because the latent dynamics model prioritizes consistent and predictable visual features learned from its training distribution. When a novel distractor appears, the model lacks an accurate latent representation of its dynamics, resulting in initial prediction errors. As predictions unfold autoregressively, these errors compound, causing the distractor's latent representation to degrade progressively. Consequently, pixels corresponding to unfamiliar distractors fail to be reliably propagated through the model's latent dynamics, leading to rapid visual distortion.

Building on this insight, we leverage a VLM's (GPT-4o) visual reasoning ability to identify problematic visual distractors through visual reasoning. Given the current observation, the world model first rolls out an in-distribution "safety-check" action plan to generate a sequence of predicted observations. The VLM is then prompted to analyze this predicted sequence. Leveraging its open-world visual reasoning capabilities, the VLM examines the temporal evolution of objects and flags those that undergo rapid distortion across frames as potential novel visual distractors. The output from the VLM is a string of object proposals deemed as novel visual distractors (shown in Figure 2).

Segmentation and inpainting. Once distractors are identified, we use a segmentation model [15] to localize and segment the corresponding regions. These regions are then passed to an image inpainting model [2], which removes the distractors and fills in the masked areas to produce a modified observation (shown in Figure 2). This intervention yields an input that more closely aligns with the world model's training distribution, alleviating hallucinations caused by artifacts from novel distractors.

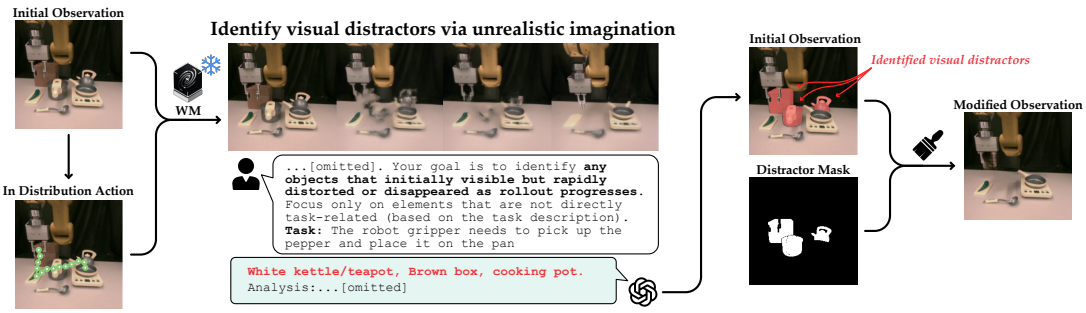

Fig. 2: **Novel visual distractor identification and observation intervention**. We leverage a VLM to analyze the temporal evolution of objects and flag those that undergo rapid distortion across frames as potential novel visual distractors (left-side; more examples and the full prompt can be found in Appendix A). These problematic distractors are inpainted to bring the observation closer to the world model's training distribution (right-side).

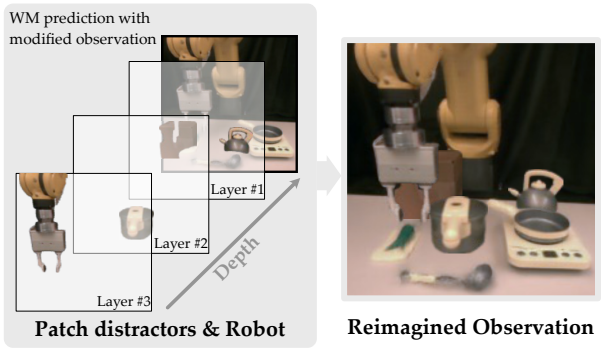

Fig. 3: **Reimagination with modified observation**. Predicted future observations from the modified input are post-processed by reintroducing previously inpainted distractors using depth-aware compositing. This ensures correct occlusion, e.g., the package box appears *behind* the gripper.

B. World Model Reimagination with Modified Observation

After intervention, we use the original world model again to predict the action plan outcomes given the modified input observation. While this mitigates hallucinated robot behaviors caused by spurious distractor artifacts, the resulting predictions no longer model the novel visual distractors since they were inpainted. To preserve visual consistency for downstream robot policy verification, we reintroduce the removed distractors into each predicted observation post-hoc.

Specifically, we first decompose each predicted observation into a set of object layers using Grounded-SAM2 [15], separating elements such as the robot, task-relevant objects, and background. For each layer, we estimate its depth from the predicted frame [13]. We also extract the inpainted distractor layer from the original observation and estimate its depth based on its original spatial placement. This distractor layer is then added back into the set of layers for every predicted frame. Finally, we reconstruct each predicted frame by compositing the layers in back-to-front depth order, ensuring that all objects are rendered with correct occlusion. This layer-wise rendering approach maintains visual realism for downstream planning, even though the distractors were absent from the rollout itself (illustrated in Figure 3).

We note that this approach may introduce physically unrealistic interactions, such as the robot gripper may appear to pass through reinserted distractors if they block the robot's motion. However, this does not compromise policy verification. Since the robot must avoid any physical contact with such distractors to ensure safety, any trajectory exhibiting these violations is automatically rejected during the verification process.

C. Action Plan Verification and Selection for Robot Visual Planning

We deploy ReOI as the visual dynamics model in the context of robot action plan verification, where a verifier needs to select action plans based on visual outcome predictions from ReOI. While our framework is agnostic to the choice of verifier, we instantiate it with a VLM that assesses each reimagined rollout in the context of the task instruction ℓ. The VLM is prompted to identify and reject plans that pose safety concerns (e.g., collisions with novel distractors or non-target objects) and to select the outcome that best aligns with the user's intent. If none of the proposed plans are deemed safe or suitable, the VLM can optionally escalate by requesting human intervention.

V. Experiment Setup

Testing environment. We conduct our evaluations using a Fanuc LR Mate 200iD/7L 6-DoF robot in a real-world robotic manipulation setup. The robot is tasked with picking and placing objects in a toy kitchen environment.

Pre-trained imitation-based generative policy. We use a Diffusion Policy [3] as the imitation-based action plan sampler. The model takes the current image observations from the wrist and third-person cameras as inputs to predict a distribution of the robot's future action plans (each action is a collection of a 3D waypoint and a gripper control signal). We use 120 multi-mode teleoperated demonstrations to train the policy.

World model training. We use the DINO-WM [27] as our base world model. DINO-WM leverages pre-trained DINOv2 [13] representation to encode visual observations and predict action outcomes directly in the DINOv2 latent representation space. We train the world model using 500 robot–environment interaction trajectories, with 200 trajectories sampled by rolling out the pre-trained diffusion policy and 300 random exploration trajectories. We separately fine-tune a DINOv2

feature decoder using images from the testing environment to map the predicted latent representations back to visual observations. More details can be found in Appendix A.

VI. Results

A. On the Effect of Novel Visual Distractors on World Model Predictions

Qualitative example. In Figure 4, we demonstrate two representative examples illustrating how novel visual distractors (highlighted in red in the first frame of each ground-truth rollout) degrade the predictive performance of the world model. Across these examples, the presence of unfamiliar distractors causes DINO-WM to hallucinate and predict that a failed action plan (one that would not successfully pick up the green pepper) would succeed (left) and mistakenly erase the target object (the green pepper) from the predicted observations (right). In contrast, ReOI produces significantly more accurate and visually consistent predictions aligned with the true task execution by reimagining future observations through test-time observation intervention.

	Full obs. eval.		In-distribution component eval.	
	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS↓
DINO-WM	0.51	0.17	0.63	0.14
ReOI	**0.94**	**0.04**	**0.79**	**0.09**

TABLE I: **Quantitative evaluations of a world model's predicted visual action outcomes**. We measure SSIM and LPIPS to evaluate the structural and perceptual similarity of the predicted visual action outcome. The left side shows results on full predicted observations, while the right side focuses on task-relevant, in-distribution objects after inpainting distractors from both the predictions and ground truth. ReOI enables the world model to produce more accurate visual predictions and effectively mitigates the hallucination of in-distribution objects caused by novel visual distractors.

Quantitative evaluation: overall prediction quality. We use two standard metrics, SSIM (Structural Similarity Index) [20] and LPIPS (Learned Perceptual Image Patch Similarity) [25], to evaluate the quality of the world model's predicted visual action outcomes. SSIM measures structural similarity based on contrast and spatial consistency, with higher scores indicating closer alignment. LPIPS evaluates perceptual similarity using deep feature comparisons, with lower scores indicating greater visual similarity. As shown in Table I, ReOI achieves better SSIM and LPIPS scores compared to DINO-WM. These improvements indicate that by shifting the input observation closer to the training distribution through test-time intervention and reimagining futures from this intervened input, our approach enables the world model to produce more accurate and robust predictions.

Quantitative evaluation: ReOI effectively mitigates the hallucination of in-distribution objects caused by novel visual distractors. Since ReOI reimagines at test time and explicitly reintroduces the novel visual distractors that DINO-WM struggles to predict, the previous evaluation does not fully disentangle the effect of these distractors on the world model's ability to predict the dynamics of in-distribution components such as the robot and target objects. To ablate this effect, we further inpaint the identified distractors in both the ground truth and predicted observations and then measure SSIM and LPIPS over the inpainted (distractor-free) predictions. The right side of Table I shows that ReOI still achieves better consistency and perceptual similarity compared to DINO-WM. These results indicate that without test-time intervention, novel visual distractors not only degrade the overall visual quality (e.g., distractors distorted across frames) but also corrupt the predicted dynamics of in-distribution objects (robot, target object), leading to hallucinated motions.

B. On the Value of Test-time Observation Intervention for Visual Planning

In this section, we evaluate the system-level performance in the context of action plan verification, where the VLM-based verifier needs to select desired action plans based on visual outcome predictions from the world model. More detailed component-level evaluation (visual distractor identification accuracy and plan verification accuracy) can be found in Appendix B).

Baselines. Our approach mitigates the impact of novel environmental variations on downstream planning by applying test-time observation intervention and reimagination. Alternatively, another intervention strategy is to reject untrustworthy world model action outcome predictions at test time to ensure safety. In addition to comparing against the base DINO-WM, we compare our approach against TrustRegion [11], which finds a region (a union of r-balls about the subset of the world model training data) where the learned visual dynamics model is deemed reliable for downstream planning and rejects world model predictions if the input observation and action plan pair falls outside of the trust region. More details about this baseline can be found in Appendix B.

Metric. We measure the task success rate and collision rate to evaluate the system-level performance of the robot visual planning. For each method, we conduct 10 trials with randomly initialized task configurations and report the average success rate. A trial is considered successful if the robot successfully and safely completes the task.

	Success Rate ↑	Collision Rate ↓
DINO-WM	0.20	0.40
TrustRegion	0.00	**0.10**
ReOI	**0.70**	**0.10**

TABLE II: **Task Success Rate**. ReOI enables more effective and safe visual planning compared to baselines.

System-level result. We present the system-level quantitative result in Table II. The results show that ReOI achieves a significantly higher task success rate compared to both DINO-WM and TrustRegion. Notably, ReOI maintains a low collision rate comparable to the conservative TrustRegion, while being substantially safer than DINO-WM that directly uses the observations for future prediction. These results suggest

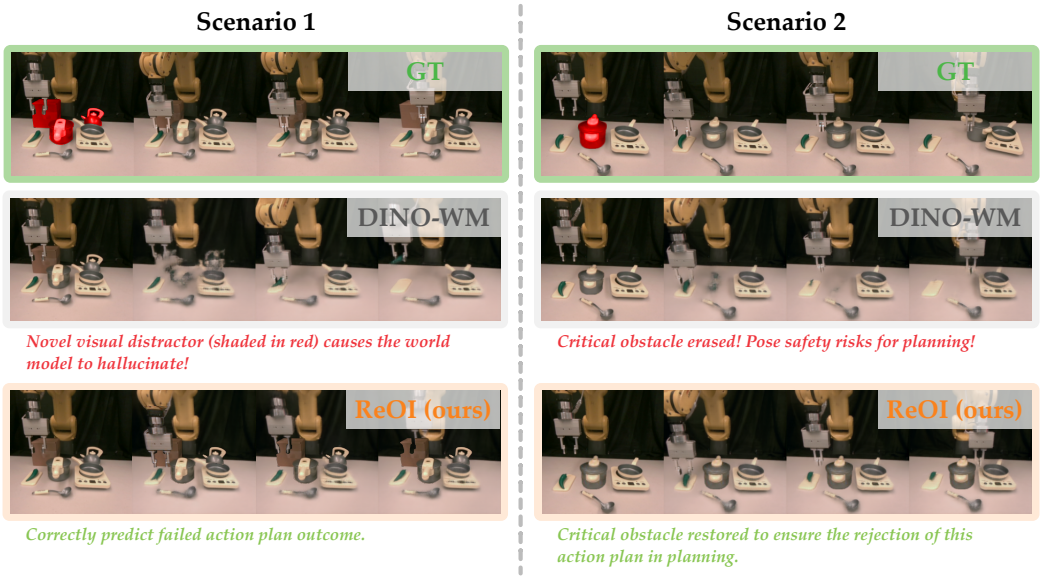

Fig. 4: **Qualitative examples of prediction quality**. The presence of unfamiliar distractor objects causes DINO-WM to hallucinate and predict that a failed action plan would succeed (left side, middle row) and mistakenly erase the target object (the green pepper) from the predicted observations (right side, middle row). **ReOI** (bottom row) effectively mitigates the effect of the novel visual distractors and produces visual outcomes more aligned with the ground truth (top row).

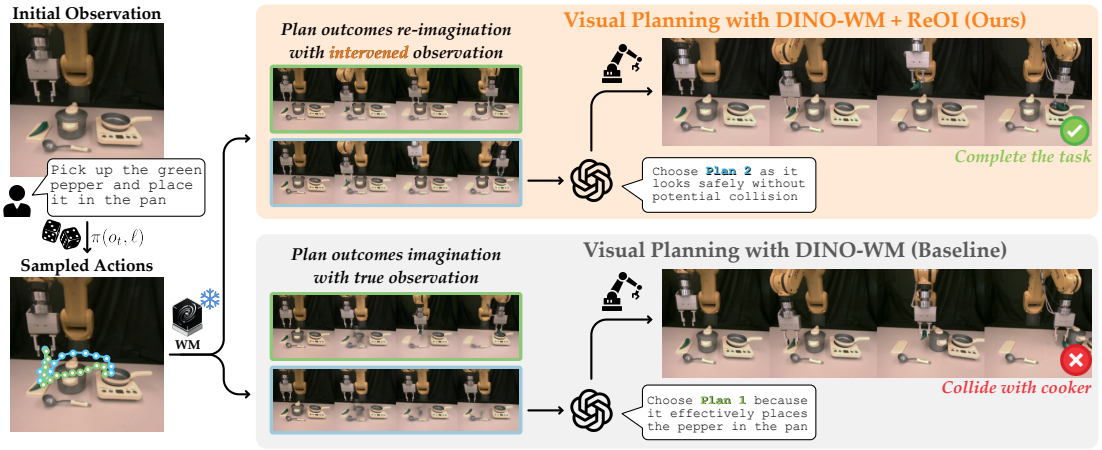

Fig. 5: **Qualitative example of robot visual planning**. The novel distractor object causes DINO-WM to hallucinate and erase a critical obstacle, leading the VLM plan verifier to incorrectly accept an unsafe action plan. In contrast, **ReOI** generates visual outcome predictions that better aligned with the reality, allowing the plan verifier to recognize the potential collision with the distractor and correctly select a safe action plan for execution.

that **ReOI** enables more effective and safe visual planning by modifying the input observation and reimagining action outcomes at test time. We show two qualitative examples in Figure 1 and Figure 5 to demonstrate the effectiveness of our approach.

VII. Conclusion

World models enable robots to "imagine" future observations given current observations and planned actions, and have been increasingly adopted as generalized dynamics models to facilitate robot learning. Despite their promise, these models remain brittle when encountering novel visual distractors. Specifically, novel distractors can corrupt action outcome predictions, causing downstream failures when robots rely on the world model imaginations for planning or action verification. In this work, we proposed Reimagination with Observation Intervention, a simple yet effective test-time strategy that enables world models to predict more reliable action outcomes in open-world scenarios where novel and unanticipated visual distractors are inevitable. We validated our approach on robotic manipulation tasks in the context of action verification, where the verifier needs to select desired action plans based on predictions from a world model. Results showed that our approach is robust to both in-distribution and unfamiliar visual distractors. Notably, it improved task success rates by up to **3×** in the presence of novel distractors, significantly outperforming action verification that relies on world model predictions without imagination interventions.

References

[1] 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.

[2] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

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

[4] Antoine Dedieu, Joseph Ortiz, Xinghua Lou, Carter Wendelken, Wolfgang Lehrach, J Swaroop Guntupalli, Miguel Lazaro-Gredilla, and Kevin Patrick Murphy. Improving transformer world models for data-efficient rl. *arXiv preprint arXiv:2502.01591*, 2025.

[5] Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *International Conference on Machine Learning*, pages 3480–3491. PMLR, 2021.

[6] Chongkai Gao, Haozhuo Zhang, Zhixuan Xu, Cai Zhehao, and Lin Shao. Flip: Flow-centric generative planning as general-purpose manipulation world model. In *The Thirteenth International Conference on Learning Representations*.

[7] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[8] Asher J Hancock, Allen Z Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. *arXiv preprint arXiv:2410.01971*, 2024.

[9] Daniel Ho, Jack Monas, Juntao Ren, and Christina Yu. 1x world model: Evaluating bits, not atoms. Technical report, 1X Technologies, Palo Alto, CA, June 2025. URL https://www.1x.tech/1x-world-model.pdf. Supplementary technical progress report available at https://www.1x.tech/discover/redwood-ai-world-model.

[10] Kaichen Huang, Shenghua Wan, Minghao Shao, Hai-Hang Sun, Le Gan, Shuai Feng, and De-Chuan Zhan. Leveraging separated world model for exploration in visually distracted environments. *Advances in Neural Information Processing Systems*, 37:82350–82374, 2024.

[11] Craig Knuth, Glen Chou, Necmiye Ozay, and Dmitry Berenson. Planning with learned dynamics: Probabilistic guarantees on safety and reachability via lipschitz constants. *IEEE Robotics and Automation Letters*, 6(3): 5129–5136, 2021.

[12] Yusuke Miyashita, Dimitris Gahtidis, Colin La, Jeremy Rabinowicz, and Jurgen Leitner. Roso: Improving robotic policy inference via synthetic observations. *arXiv preprint arXiv:2311.16680*, 2023.

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[14] Han Qi, Haocheng Yin, Yilun Du, and Heng Yang. Strengthening generative robot policies through predictive world modeling. *arXiv preprint arXiv:2502.00622*, 2025.

[15] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[16] Martin Riedmiller, Tim Hertweck, and Roland Hafner. Less is more–the dispatcher/executor principle for multi-task reinforcement learning. *arXiv preprint arXiv:2312.09120*, 2023.

[17] Shenghua Wan, Yucen Wang, Minghao Shao, Ruying Chen, and De-Chuan Zhan. Semail: eliminating distractors in visual imitation via separated models. In *International Conference on Machine Learning*, pages 35426–35443. PMLR, 2023.

[18] Tongzhou Wang, Simon S Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised mdps: Learning world models better than the world itself. *arXiv preprint arXiv:2206.15477*, 2022.

[19] Yucen Wang, Shenghua Wan, Le Gan, Shuai Feng, and De-Chuan Zhan. Ad3: Implicit action is the key for world models to distinguish the diverse visual distractors. In *International Conference on Machine Learning*, pages 51546–51568. PMLR, 2024.

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[21] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.

[22] Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. *arXiv preprint arXiv:2502.01828*, 2025.

[23] Jiange Yang, Wenhui Tan, Chuhao Jin, Keling Yao, Bei Liu, Jianlong Fu, Ruihua Song, Gangshan Wu, and Limin Wang. Transferring foundation models for generalizable robotic manipulation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1999–2010. IEEE, 2025.

[24] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representa-

tions for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[26] Haoyu Zhen, Qiao Sun, Pengxiao Han, Siyuan Zhou, Yilun Du, and Chuang Gan. Learning 4d embodied world models.

[27] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[28] Chuning Zhu, Max Simchowitz, Siri Gadipudi, and Abhishek Gupta. Repo: Resilient model-based reinforcement learning by regularizing posterior predictability. *Advances in Neural Information Processing Systems*, 36: 32445–32467, 2023.