Enhancing Visuomotor Manipulation Control with a Predictive-Dynamics-Aware World-Model-Conditioned Diffusion Policy

Sangho An[1†], Geunha Kim[2†], and Jun Moon[1*]

[1]Department of Electrical Engineering, [2]Department of Artificial Intelligence Systems

Hanyang University
Seoul, Korea

The 26th International Conference on Control, Automation and Systems (ICCAS 2026)
Paper WeBT8.3 · Wednesday, October 28, 2026, 16:10–16:25

[†]Equal contribution, [*]Corresponding author

Outline

Introduction

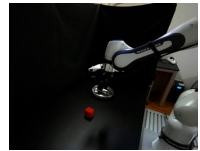
Methodology

Experiment Results

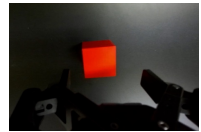
Conclusion

Introduction: Research Background

- Research area: Visuomotor policies map camera images directly to robot actions [1–3].
- Diffusion Policy [8]: Generates an action *sequence* by iterative denoising ⇒ handles multi-modal demonstrations better than plain behavior cloning.
- Technical difficulty: The same image can require *different* actions, depending on object pose, contact timing, and the robot's previous motion.
- Core message: DP is conditioned mainly on *image features of 2 frames*; it does not explicitly encode *how the scene is evolving*.



Side camera



Wrist camera

A single snapshot cannot tell
"Am I approaching, grasping, or lifting?"
⇒ need a **dynamics-aware** cue from recent history

Introduction: Research Goal

- Objective: Improve real-world visuomotor manipulation of diffusion policy by injecting *temporal / dynamics* information.
- Key idea: A lightweight **GRU world model** summarizes the recent observation history into a cue h_t^{wm}, which is simply **concatenated** to the DP condition.
- Contributions:
  - (i) **WM-DP**: A simple world-model conditioning for diffusion policy (one concatenation).
  - (ii) **WM outside the denoising loop**: One WM pass per action chunk, only **+70 ms**.
  - (iii) **Real-robot validation** on a Franka Research 3: Red block 45.0% $\rightarrow$ **72.5%**, seen HYlion poses 42.5% $\rightarrow$ **68.8%**; limitation on unseen 90° rotations.

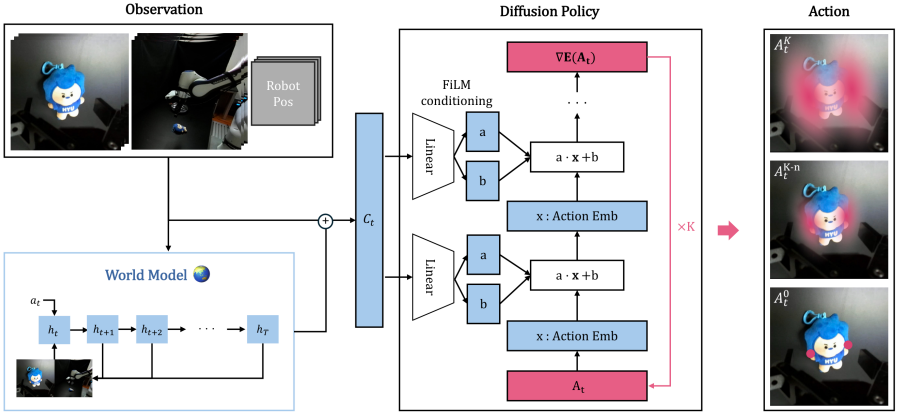
Research question: Can a *simple, early-stage* world-model conditioning improve diffusion policy on a real robot?

Introduction: Related Works

Approach	Method / assumption	Limitation for our setting
Diffusion Policy [8]	Denoises action sequences conditioned on recent observations (CNN image features)	No explicit representation of scene dynamics
World models / RSSM [15,18–21]	Learn latent dynamics; used for planning and RL by *imagined rollouts*	Heavy: stochastic latents, reconstruction, rollouts
DiWA [24]	World model as a learned *simulator* to fine-tune DP offline with RL	Extra RL stage; WM not a conditioning signal
WM-DP (ours)	WM latent used *only as an extra condition* of the denoising network	No rollout, no RL, one WM pass per action chunk

- Research gap: Whether a *minimal* world-model *conditioning* branch (not a simulator) already helps diffusion policy on a *real* robot remains unexplored.

Methodology: Overall Architecture

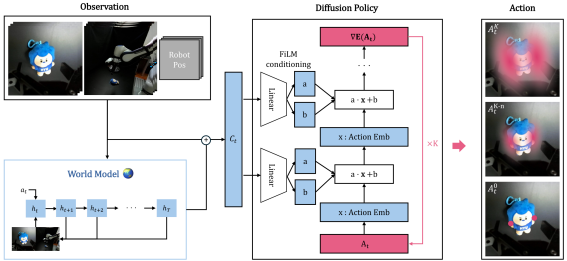


Input: Side + wrist RGB (224×224) and robot joint state (incl. gripper)

World model: GRU over the last **8** observations → h_t^{wm}

Output: Concat. condition $\tilde{C}_t$ → FiLM U-Net denoises action chunk A_t

Methodology: Problem Formulation



World model + **conditioning** = the only change to DP

Condition

$$\tilde{C}_t = \left[\, O_t^{\text{obs}} \,;\, h_t^{\text{wm}} \,\right]$$

FiLM

$$\text{FiLM}(x) = \gamma(\tilde{C}_t, k) \odot x + \beta(\tilde{C}_t, k)$$

Loss

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda \left\| h_t^{\text{wm}} - a^* \right\|^2$$

$$\mathcal{L}_{\text{diff}} = \mathbb{E} \left\| \epsilon - \epsilon_\theta(A_t^k, k, \tilde{C}_t) \right\|^2$$

$$\lambda = 0.1$$

Methodology: Algorithm (Closed-Loop Inference)

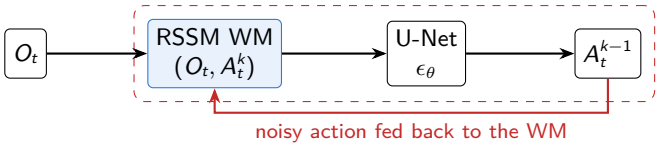
Algorithm Outline (one planning cycle, robot control at 10 Hz)

1. Keep a rolling window of the last 8 observations $O_{t-7:t}$ (images + joint state).
2. Encode the last 2 frames $\rightarrow O_t^{\text{obs}}$; run the GRU world model **once** $\rightarrow h_t^{\text{wm}}$.
3. Sample $A_t^K \sim \mathcal{N}(0, I)$ and denoise for $K = 80$ DDPM steps:
   $A_t^{k-1} = \text{Step}\big(A_t^k,\, \epsilon_\theta(A_t^k, k, \tilde{C}_t)\big)$ – the condition $\tilde{C}_t$ is **fixed** during denoising.
4. Execute the first 8 of the 16 predicted actions (0.8 s), then replan (receding horizon).

- Design parameters: WM horizon 8, obs. horizon 2, prediction / execution horizon 16 / 8, GRU hidden 256.
- Main novelty: The world model is a *conditioning branch*, not a simulator – it never sees the noisy actions, so its cost is paid **once per action chunk**.

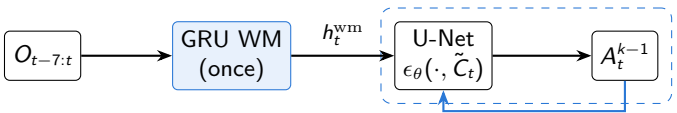
Methodology: Why a Lightweight GRU, not an RSSM in the Loop?

(a) First attempt: RSSM / action-conditioned WM



cost $\approx K\,(c_{\mathrm{WM}} + c_\theta)$
WM re-run at every denoising step

(b) WM-DP (ours): history-only GRU WM

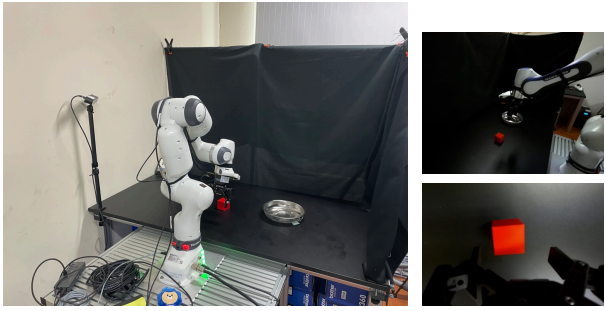


cost $\approx c_{\mathrm{WM}} + K\,c_\theta$
measured: +70 ms vs. DP (K = 80)

Design principle

We first explored RSSM (Dreamer-style) world models with imagination. Their condition depends on the noisy action, so the WM must be re-evaluated inside the K-step denoising loop $\Rightarrow$ too slow for 10 Hz control. Decoupling the WM from the denoising loop keeps the overhead to a **single** forward pass.

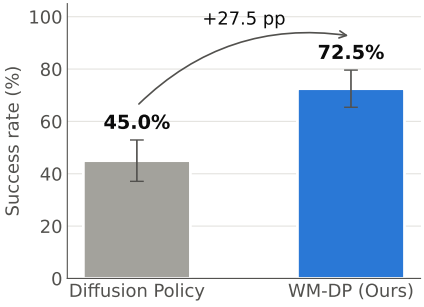
Experiment: Setup



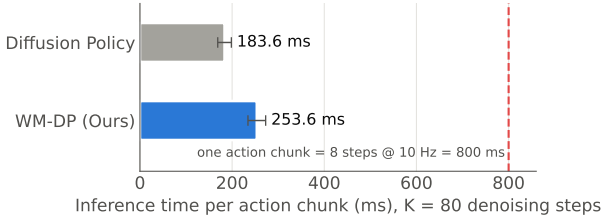
Franka Research 3 + Robotiq gripper; ZED 2 (side view) + ZED Mini (wrist view)

- Scenario: Grasp → lift over a *hurdle* → place in basket → release *only after contact*.
- Objects: **Red block** (rigid, slips unless aligned) and **HYlion** doll (deformable, orientation-dependent look).
- Data: 72 teleoperated demos (36 per object).
- Parameters: 75 epochs, obs. horizon 2, WM horizon 8, $K = 80$.
- Baseline: Original DP with the *same* data and settings.
- Protocol: 40 trials per condition (10 trials $\times$ 4 seeds).

Experiment: Red Block – Success Rate and Inference Time

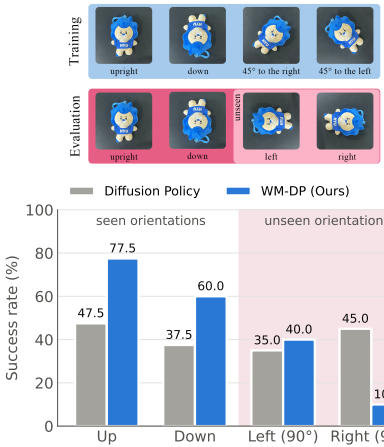


Success rate, mean ± standard error (40 trials each)



- Success: **45.0%** → **72.5%** (+27.5 pp).
- Cost: Inference **+70 ms**, well within one 0.8 s action chunk.
- Behavior: DP grasps too shallowly, slips, or releases early; WM-DP grasps deeper, re-grasps, and holds until contact.

Experiment: Object-Centric Grasping with HYlion



Top: training / evaluation orientations. Bottom: success rate (%), 40 trials each

	DP	WM-DP
Seen (Up, Down)	42.5	**68.8**
Unseen (Left, Right)	**40.0**	25.0

- Comparison: Better in Up / Down / Left (+30.0 / +22.5 / +5.0 pp).
- Failure case: Right 90°: 45.0 → 10.0% – demos never rotate the end-effector beyond 45°.
- Discussion: The WM cue is *actively used*; it helps inside the demonstrated dynamics range, but does not extrapolate beyond it.

Conclusion

- Summary: WM-DP conditions diffusion policy on the latent cue of a lightweight GRU world model computed from the recent observation history – a single concatenation.
- Main contribution: Red-block success **45.0%** $\rightarrow$ **72.5%** with only **+70 ms** inference, with *no* extra data and the same training budget.
- Validation: Deeper and more stable grasps, re-grasping, and contact-aware release on a real Franka; gains on seen HYlion orientations (42.5% $\rightarrow$ 68.8%).
- Future work: Generalization beyond the demonstrated dynamics (unseen 90° rotations), richer world models (e.g., RSSM latents) kept *outside* the denoising loop, and faster samplers.

Thank you!

Questions & Comments



Experiment video: youtu.be/dVcs-mAfopQ

Sangho An · shan1005@hanyang.ac.kr Geunha Kim · rlarmsgk2@hanyang.ac.kr

Backup: Qualitative Behavior (Red Block)

Diffusion Policy: failure patterns	WM-DP: observed behavior
Approaches the block, but the grasp is **too shallow** → blocked by the hurdle	**Deeper, more stable** grasps, also in harder workspace regions
Misaligned approach → block slips or is pushed aside while closing	**Re-grasping**: adjusts the grasp before lifting
Early gripper opening → drops the object during transport	Holds until the object **contacts** the basket bottom

- Key message: The history cue helps the policy reason about the *phase* of the interaction (approach / grasp / transport / release), not only about the current image.

Experiment video

youtu.be/dVcs-mAfopQ

Backup: Implementation Details

	Diffusion Policy	WM-DP (ours)
Image encoder	ResNet-18 per camera (GroupNorm), 224×224	
Condition dimension	2064	$2064 + 8 = 2072$
World model	–	ResNet-18$\times 2$ + GRU(256) + MLP
Observation / WM horizon	2 / –	2 / 8
Prediction / execution horizon	16 / 8 (7 joint positions + gripper, 10 Hz)	
Diffusion	DDPM, 100 training steps, $K = 80$ sampling steps	
Training	75 epochs, AdamW (lr 10^{-4}), EMA 0.995	
Extra loss	–	$\lambda \lVert h^{\mathrm{wm}} - a^{*} \rVert^2$, $\lambda = 0.1$
Parameters	49.2 M	72.6 M
Inference time	183.6 ± 14.8 ms	253.6 ± 19.2 ms

Backup: Statistical View of the Results

Condition (40 trials each)	DP	WM-DP	Fisher exact p
Red block	18 / 40 (45.0%)	29 / 40 (72.5%)	0.023
HYlion Up	19 / 40 (47.5%)	31 / 40 (77.5%)	0.011
HYlion Down	15 / 40 (37.5%)	24 / 40 (60.0%)	0.073
HYlion Left (unseen)	14 / 40 (35.0%)	16 / 40 (40.0%)	0.82
HYlion Right (unseen)	18 / 40 (45.0%)	4 / 40 (10.0%)	0.001
HYlion seen (Up + Down)	34 / 80 (42.5%)	55 / 80 (68.8%)	0.001

- Trial counts follow from the reported standard errors ($\mathrm{SE} = \sqrt{p(1-p)/n}$, $n = 40$).
- Post-hoc test (not in the paper): gains are significant on the red block and on seen HYlion orientations (pooled); the Right-90° drop is also significant.

Backup: Design Evolution of the World-Model Branch

1. Action-conditioned latent WM (simulation, robomimic *Can*): Predicts future visual latents from images + a candidate action sequence → needs the (noisy) action → evaluated inside the denoising loop.

2. Dreamer-style RSSM (discrete stochastic latent, KL balancing): With imagination conditioning and dream distillation over multiple candidate action chunks → repeated sampling + rollouts per decision.

3. Final WM-DP: History-only GRU world model, **one** forward pass per action chunk, trained jointly with a light auxiliary loss → real-time on the Franka at 10 Hz.

Take-away

For real-robot diffusion policies, *where* the world model sits relative to the K-step denoising loop matters as much as *how expressive* it is.

Backup: Limitations and Next Steps

- Unseen rotations: Demos contain end-effector rotations up to $\sim 45^\circ$; the 90° cases require out-of-distribution wrist motion. *Hypothesis:* the learned cue biases the policy toward in-distribution grasp poses (Right 90°: 10%).
- Confounders to ablate: Longer history alone (DP with 8-frame observation), parameter count (49.2 M vs. 72.6 M), and WM-only conditioning.
- Scope: Two objects, one task family, single robot; 40 trials per condition.
- Next: (i) Rotation-augmented demonstrations, (ii) richer latent dynamics (RSSM state) kept outside the loop, (iii) fewer-step samplers (e.g., DDIM) to cut latency for both methods.

Backup: Training Data Composition

Object	Orientation in demos	Demos
Red block	Various positions	36
HYlion	Upright	Most of 36
	Inverted (down)	Very few
	45° left / 45° right	Very few
Total (one single policy, both objects)		**72**
Eval. only	90° left / 90° right	0 (unseen)



Top: training poses. Bottom: evaluation poses.

- One dataset, one model: All 72 demos of both objects train a *single* policy.
- No data augmentation: Images are only resized (224×224); no crop, color, or rotation augmentation.
- Take-away: Even with rare non-upright demos, WM-DP reached 60.0% on the inverted pose (DP: 37.5%) $\Rightarrow$ gains come from limited, imbalanced data, not from extra data or augmentation.