Supplementary Ablation Study on the World Synesthesia Model

Anonymous Author(s)
Affiliation
Address
`email`

1 Additional World Synesthesia Model Ablations

The main paper shows that the World Synesthesia Model (WSM) improves dexterous in-hand rotation beyond raw visuotactile observation fusion. This supplementary study adds **controlled WSM variants that are not included in the main manuscript** and separates three possible sources of improvement: **clean geometric supervision**, **recurrent latent dynamics**, and **a learned policy feature**. All variants are evaluated in the same z-axis multi-object benchmark as the main WSM analysis, with the same reward, randomized evaluation protocol, deployable policy interface, and object set.

Evaluation design. The first variant replaces the clean-depth reconstruction target with a **noisy-depth target** while preserving the same noisy encoder input. This control tests whether the performance gain comes specifically from denoising supervision rather than from a generic depth reconstruction objective. The second variant weakens recurrent state propagation by **removing the previous deterministic state** h_{t-1} from the recurrent update, while retaining the RSSM-style WSM branch and its downstream policy feature. This evaluates the role of deterministic memory in the action-conditioned latent dynamics. The third variant removes the recurrent RSSM branch and uses a **per-step encoder–decoder latent** z_t as the policy context. This encoder-decoder control preserves a learned perceptual feature but removes temporal dynamics, the dynamics prior, and KL regularization.

We keep other training-only auxiliary heads, such as object-state-, value-, and shape-related predictions, fixed across these controls. These heads follow common practice in asymmetric robot-control representation learning, are not used by the deployable actor at test time, and are not claimed as our contribution. In preliminary experiments, they brought only very marginal gains and did not change the qualitative conclusion. We therefore focus this ablation on the two mechanisms central to WM-Craftnet: **clean-depth supervision** and **action-conditioned recurrent inference**.

Table 1: Additional WSM ablations on the z-axis multi-object benchmark. Metrics are reported as mean $\pm$ 95% confidence interval under the same evaluation protocol as the main table.

Method	EpLen↑	Return↑	RotR↑	OffAxis↓	AngVar↓
WM-Craftnet (noisy depth supervision)	431.1 ± 2.0	*708.0 ± 3.5*	1.265 ± 0.005	1.299 ± 0.006	1.456 ± 0.045
WM-Craftnet (no h_{t-1} recurrent input)	*434.3 ± 2.3*	705.4 ± 4.3	*1.282 ± 0.005*	**1.125 ± 0.006**	**1.185 ± 0.008**
WM-Craftnet (encoder-decoder z policy)	417.4 ± 2.6	667.6 ± 4.7	1.254 ± 0.003	1.323 ± 0.006	1.343 ± 0.018
WM-Craftnet (full WSM)	**435.2 ± 1.9**	**753.3 ± 3.6**	**1.293 ± 0.004**	*1.225 ± 0.006*	*1.324 ± 0.027*

Results. Figure 1 shows that all three controls train below the full WM-Craftnet curve. Table 1 confirms the same trend at evaluation time: the **full WSM reaches the highest return** (753.3) and **highest rotation rate** (1.293), while maintaining a long episode length. The encoder-decoder z-policy remains stable but has lower return and episode length, indicating that **a per-frame perceptual feature is insufficient to match the full recurrent WSM**.

Noisy-depth supervision reaches 708.0 return, substantially below the full clean-depth WSM. This supports the interpretation that the clean-depth target is not merely an auxiliary reconstruction task; rather, it regularizes the recurrent state toward a denoised hand-object geometry that is useful for control under corrupted wrist-depth observations. The **no-h_{t-1} recurrent-input variant** achieves favorable off-axis motion and angular-velocity variation, but its lower return suggests that weakening deterministic memory propagation reduces the value of the WSM as a compact interaction history.

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

Table 2: Real-robot z-axis rotation with an additional clear-depth IHR control. RR reports radians of rotation in a 20-s episode, and SR reports the success rate over 10 trials.

Method	Duck		Cross block		Corner block		Double-notched block(unseen)	
	RR	SR	RR	SR	RR	SR	RR	SR
Open-loop Replay	2.32	6/10	*3.04*	6/10	*5.19*	*7/10*	0.23	0/10
Touch Dexterity[1]	1.54	4/10	0.34	2/10	0.22	0/10	0.27	0/10
In-Hand Rotation[2]	1.83	5/10	0.45	1/10	0.36	1/10	0.39	0/10
In-Hand Rotation with clear depth	*2.76*	*8/10*	1.21	*8/10*	2.43	**10/10**	*0.80*	0/10
WM-Craftnet	**16.179**	**10/10**	**8.01**	**10/10**	**8.48**	**10/10**	**4.32**	**8/10**

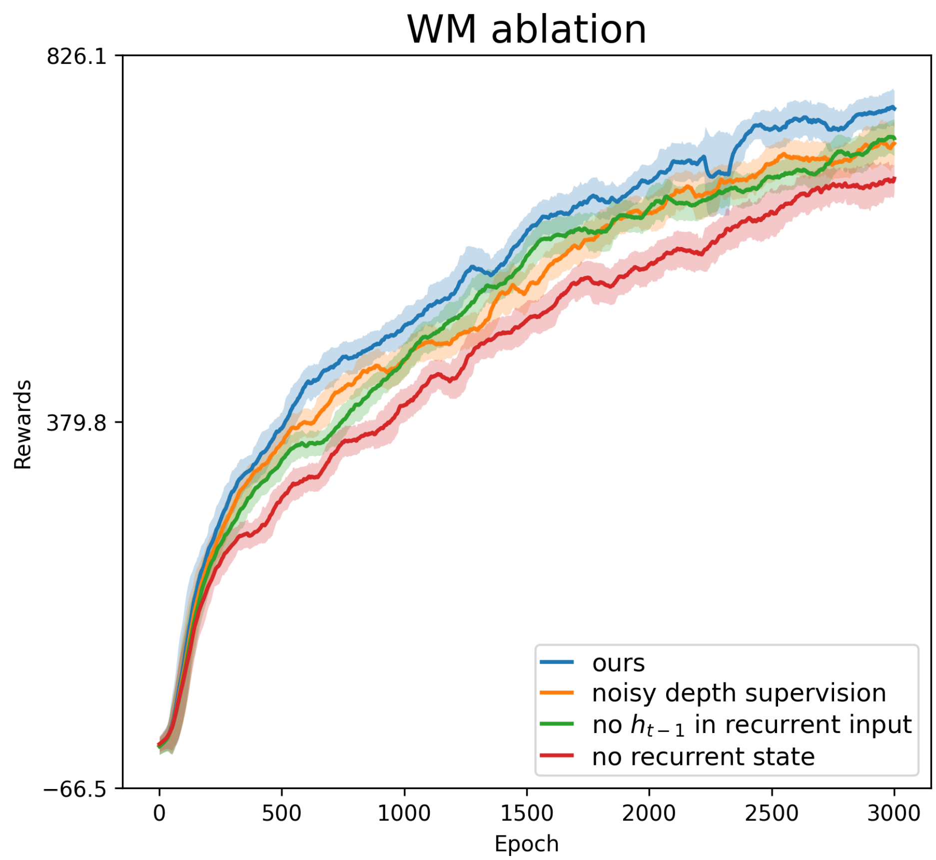

Figure 1: Training curves for additional WSM controls on the z-axis multi-object benchmark. The full WM-Craftnet model learns faster and reaches a higher final reward than variants with noisy-depth supervision, weakened recurrent-state input, or no recurrent state.

2 Real-World Clear-Depth Control

We further evaluate whether the sim-to-real performance gain can be explained only by providing cleaner depth to an existing visuotactile policy. Here, *clear depth* refers to WSM-denoised depth predicted from real noisy wrist-depth observations, not ground-truth or privileged depth. Table 2 adds a **new IHR with clear depth control** to the real-robot z-axis evaluation. This experiment is distinct from the main manuscript because it tests whether the observed hardware advantage is attributable to **depth denoising alone**. The project website provides synchronized videos of real noisy wrist depth, WSM-predicted depth, continuous multi-object rollouts, and perturbation recoveries.[1]

[1]Project website: `https://wmcraftnet.github.io`.

WSM-denoised depth substantially improves IHR over the noisy-depth setting, especially on the cross block and corner block, confirming that depth corruption is an important sim-to-real bottleneck. However, the IHR policy with clear depth remains far below WM-Craftnet in rotation amount and fails on the unseen double-notched block. This gap indicates that the real-world advantage of WM-Craftnet is **not explained by cleaner depth alone**. The WSM-predicted state provides a denoised and temporally informed representation of hand-object geometry and contact evolution, which is consistent with the qualitative depth-reconstruction and recovery videos shown on the project website.

Conclusion. Taken together, the additional simulation controls and real-world clear-depth comparison support the same mechanism: WM-Craftnet benefits from the **combination of clean geometric supervision and action-conditioned recurrent inference**. Denoising perception is useful, but the full WSM is more effective because it couples denoised geometry with temporal contact and motion information needed for robust in-hand manipulation.

References

[1] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang. Rotating without seeing: Towards in-hand dexterity through touch. In *Robotics: Science and Systems (RSS)*, 2023. URL https://arxiv.org/abs/2303.10880.

[2] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565. IEEE, 2024. URL https://arxiv.org/abs/2312.01853.