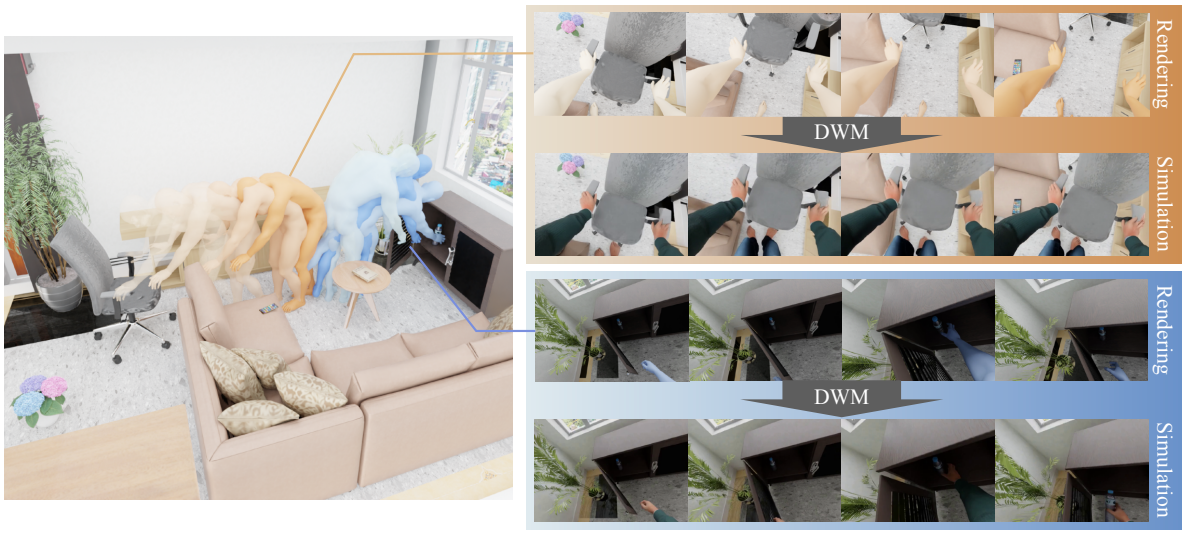
Dexterous World Models

Byungjun Kim[1*] Taeksoo Kim[1*] Junyoung Lee[1] Hanbyul Joo[1,2]

[1]Seoul National University [2]RLWRLD

https://snuvclab.github.io/dwm/



Embodied Actions in Static 3D Scenes Egocentric Simulation from Dexterous World Models

Figure 1. **Dexterous World Models** predict egocentric visual dynamics of static 3D scenes, driven by dexterous hand manipulations.

Abstract

*Static 3D reconstruction has made it increasingly practical to build realistic digital twins of everyday environments, but these reconstructions remain largely non-interactive: they support navigation and view synthesis, yet cannot predict how dexterous actions change the world. We introduce **Dexterous World Models (DWM)**, a scene-action-conditioned video diffusion framework that simulates egocentric visual dynamics induced by human hand manipulation in a known static 3D scene. Given a rendering of the static scene along a camera trajectory and an egocentric hand-mesh video encoding the action, DWM generates a temporally coherent interaction video while preserving unaltered regions of the scene. The model is initialized from a video inpainting diffusion prior, encouraging identity preservation for static content and generative residual modeling for action-induced changes. Because no large-scale real dataset provides aligned static-scene, hand-action, and interaction triplets under moving egocentric cameras, we train with a hybrid dataset that combines aligned synthetic interactions with fixed-camera real-world videos. Experiments show that DWM produces plausible object motion and articulation while preserving scene consistency, marking a step toward interactive digital twins driven by dexterous egocentric actions.*

1. Introduction

World models [6] are predictive models of environment dynamics: given the current world state and an action, they estimate what will happen next. For embodied agents and human-centered digital twins, the relevant dynamics are often caused not by camera motion alone, but by dexterous interaction: hands grasp objects, open articulated parts, displace tools, trigger mechanisms, and otherwise change

*Equal contribution

the state of the environment. Modern reconstruction methods [14, 15] make it easy to obtain visually faithful static 3D scenes, yet these digital twins typically remain inert. They can be rendered from novel viewpoints, but they do not explain how the scene should evolve when a person acts inside it.

Recent video generative models provide a promising route to visual world modeling, but common formulations entangle three different factors: the static scene, the camera trajectory, and the action-induced state change. Image-to-video systems hallucinate both the environment and the dynamics from a partial observation, while navigation-oriented world models primarily treat camera motion as the action [3, 20]. This is insufficient for everyday interaction. Camera motion changes what is observed; dexterous hand motion changes the world itself. Text prompts are also a weak action representation because they cannot precisely specify hand pose, timing, contact geometry, or target selection. As a result, a model that synthesizes the entire video from text and an initial image can alter the background, miss the intended object, or produce motion that is semantically plausible but physically ungrounded.

We propose **Dexterous World Models (DWM)**, a formulation and implementation for egocentric visual simulation in static 3D scenes. Instead of asking a video model to regenerate the whole scene, DWM receives the static scene as input and learns to synthesize the residual visual dynamics caused by dexterous hand actions. The input consists of two temporally aligned egocentric videos: a static-scene rendering along the target camera trajectory, and a hand-mesh rendering that specifies the manipulation. The output is an interaction video that preserves the static environment wherever no action-induced change occurs while generating plausible object dynamics near the hands. This separation is central to the method: the static scene provides spatial and appearance consistency, while the hand trajectory provides fine-grained action control.

Our implementation instantiates DWM as a scene-action-conditioned latent video diffusion model initialized from an inpainting prior. A full-mask inpainting model is already trained to reconstruct known content while relying on a learned generative prior for uncertain regions. We reuse this behavior as an identity-preserving initialization: the static-scene video anchors the output, and hand-motion conditioning guides the model to create only the manipulation-induced residuals. To train the model, we construct hybrid supervision from synthetic and real videos. Synthetic human-scene interactions from TRUMANS [7] provide exactly aligned triplets under moving egocentric viewpoints, while fixed-camera real-world interaction videos from TASTE-Rob [19] provide realistic object dynamics and visual variation.

The resulting model brings interactivity to static digital twins without requiring explicit object-level articulation models or manually specified physical simulators. Across synthetic and real-world benchmarks, DWM outperforms strong video-editing and interaction-generation baselines, including text-guided SDEdit [9], a fine-tuned inpainting model [1], and InterDyn [2] in the static-camera setting. Qualitatively, it generates coherent manipulations such as grasping, moving objects, and opening articulated structures while maintaining camera and scene consistency.

Our contributions are: (1) a dexterous world-modeling formulation that separates static scene rendering from action-induced dynamics; (2) a video diffusion architecture conditioned on aligned static-scene and egocentric hand-mesh videos; (3) a hybrid training strategy combining aligned synthetic triplets with fixed-camera real-world interactions; and (4) experiments showing realistic egocentric interaction simulation and generalization to unseen real-world dynamic viewpoints.

2. Method

Dexterous World Modeling. Let $\mathbf{S}_0$ denote a known static scene, $\mathcal{C}_{1:F}$ an egocentric camera trajectory, and $\mathcal{H}_{1:F}$ a dexterous hand trajectory. The goal is to predict the interaction video $\mathbf{V}_{1:F}$ that would be observed after applying the hand action in the scene. We view the future world as $\mathbf{S}_0 + \Delta\mathbf{S}_{1:F}$, where $\Delta\mathbf{S}_{1:F}$ represents the residual state change induced by the action. Unlike image-to-video world models that must infer both the static scene and its dynamics from an initial frame, DWM conditions directly on $\mathbf{S}_0$ and learns only the action-dependent residual:

$$p_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \mathcal{C}_{1:F}, \mathcal{H}_{1:F}). \tag{1}$$

The camera trajectory determines how the evolving world is observed, while hand motion determines how the world changes. This factorization reflects the causal distinction between navigation and manipulation: moving the camera changes viewpoint, but grasping or pushing changes scene state.

Egocentric Conditioning. We implement the formulation through egocentric renderings. Given a rendering operator $\Pi(\cdot; \mathcal{C}_{1:F})$, DWM receives the static-scene video $\Pi(\mathbf{S}_0; \mathcal{C}_{1:F})$ and the hand-mesh video $\Pi(\mathcal{H}_{1:F}; \mathcal{C}_{1:F})$. The static video carries geometry, appearance, and camera motion, while the hand video provides pixel-aligned action cues, including hand location, articulation, and temporal contact structure. The model approximates

$$p(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \mathcal{C}_{1:F}, \mathcal{H}_{1:F}, \mathcal{T}) \approx p(\mathbf{V}_{1:F} \mid \Pi(\mathbf{S}_0; \mathcal{C}_{1:F}), \Pi(\mathcal{H}_{1:F}; \mathcal{C}_{1:F}), \mathcal{T}), \tag{2}$$

where $\mathcal{T}$ is a text prompt used as semantic guidance. This representation is especially useful for egocentric interaction because the camera naturally follows the actor's attention,

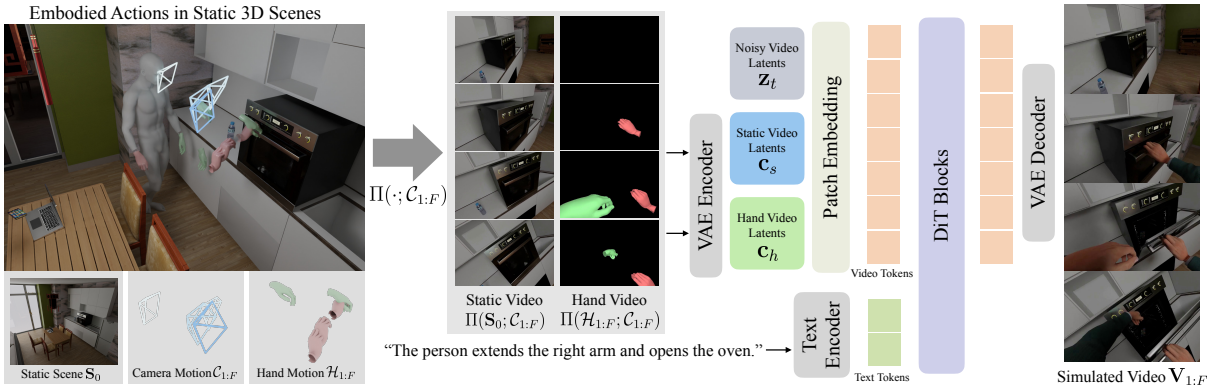


Figure 2. **Overview.** DWM simulates egocentric visual dynamics induced by embodied actions within a given static 3D scene. We instantiate it as a video diffusion model conditioned on the egocentric projections of the static scene and hand trajectories.

and the projected hand motion is spatially aligned with the scene locations where dynamics should occur.

Video Diffusion Architecture. DWM is a latent video diffusion model built on a pretrained video VAE and Diffusion Transformer (DiT) [12]. The target interaction video, static-scene video, and hand-mesh video are encoded into latent tensors. At diffusion step t, the DiT predicts the noise in the noised target latent $\mathbf{z}_t$ conditioned on the static and hand latents $(\mathbf{c}_s, \mathbf{c}_h)$:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{z}_0, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t \mid \mathbf{c}_s, \mathbf{c}_h)\|_2^2\right]. \quad (3)$$

The conditional latents are concatenated channel-wise with the noisy video latent before being processed by the transformer. At inference time, iterative denoising produces a latent interaction video that is decoded into RGB frames.

Residual Dynamics via Inpainting Priors. A key design choice is to initialize DWM from a pretrained video inpainting diffusion model [1]. When all pixels are marked as known, an inpainting model behaves like an approximate identity function, reconstructing the input while preserving spatial structure and temporal coherence. We exploit this behavior for residual dynamics learning. The static-scene video should remain unchanged except where the hand action causes visible effects:

$$\mathbf{V}_{1:F} = \Pi(\mathbf{S}_0; \mathcal{C}_{1:F}) + \Delta\mathbf{V}_{1:F}. \quad (4)$$

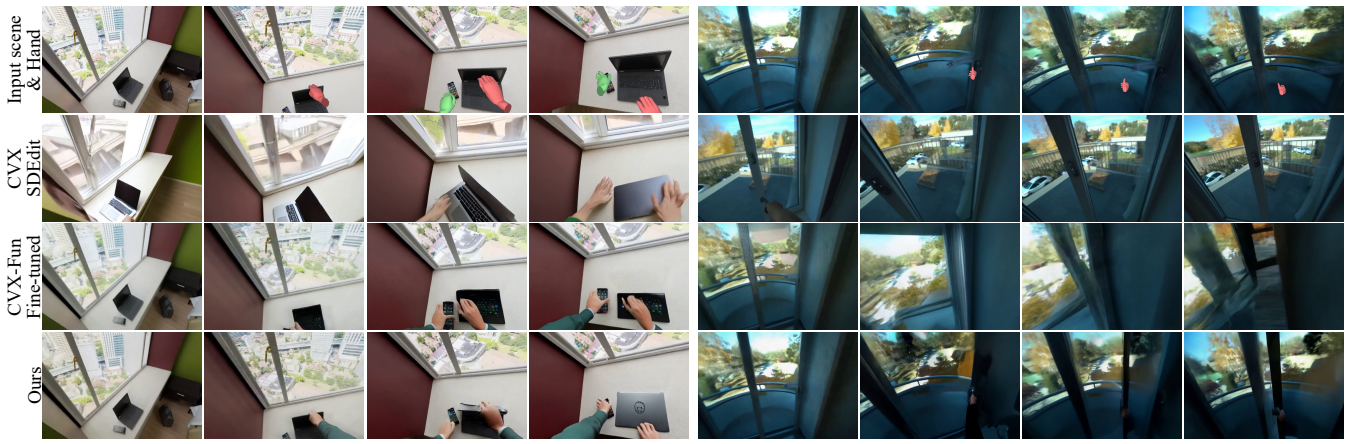
Initializing from an inpainting prior biases the model toward copying the static input and using its generative capacity to synthesize $\Delta\mathbf{V}_{1:F}$, such as object displacement, articulation, or appearance changes caused by contact. This reduces unnecessary background hallucination and stabilizes training compared with generating the entire video from scratch.

Hybrid Dataset Construction. Training requires triplets of static-scene videos, hand-action videos, and corresponding interaction videos under aligned camera trajectories. Such triplets are difficult to collect at scale in real environments because one would need both the pre-action static scene and the dynamic interaction observed from the same moving egocentric camera. We therefore combine two complementary data sources. First, TRUMANS [7] provides synthetic 3D human-scene interactions with controllable scene state and camera motion. For each sequence, we render synchronized egocentric outputs: the interaction video, the static scene replayed under the same camera trajectory, and the hand mesh segmented from the SMPL-X body model [10, 13]. Second, fixed-camera real-world videos from TASTE-Rob [19] provide realistic physical dynamics. Because the camera is static, the first frame can be repeated to form an aligned static-scene video, and hand meshes are estimated with HaMeR [11].

For evaluation under real dynamic egocentric viewpoints, we collect 60 Aria Glasses sequences with SLAM camera trajectories [4]. The operator first scans the static scene and then performs an interaction. Pre-action frames are reconstructed as a 3D Gaussian scene [8, 17], which is rendered along the recorded interaction trajectory to obtain a static-scene video aligned with the ground-truth interaction. This protocol covers diverse interactions, including pick-and-place, articulated object manipulation, and counterfactual effects, allowing us to measure whether DWM generalizes beyond fixed-camera real training data.

3. Experiments

Benchmark. We evaluate on 144 held-out triplets of static-scene video, hand-mesh video, and ground-truth interaction video. The benchmark covers synthetic dynamic-camera TRUMANS sequences, unseen real-world static-camera TASTE-Rob videos, and Aria-captured real-world dynamic-camera samples with reconstructed static scenes. We report



(a) Unseen synthetic scene (b) Unseen real-world scene

Figure 3. **Qualitative comparison under dynamic egocentric views.** DWM generates physically plausible interactions following the input hand actions and generalizes to unseen real-world dynamic-camera scenes, such as opening a sliding window.

	Synthetic				Real-World							
	Dynamic Camera				Static Camera				Dynamic Camera			
	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
CVX SDEdit	19.424	0.675	0.464	0.257	16.194	0.586	0.446	0.224	19.154	0.507	0.676	0.492
CVX-Fun Fine-tuned	20.541	0.767	0.370	0.175	18.951	0.780	0.265	0.089	18.129	0.472	0.591	0.328
InterDyn	–	–	–	–	19.331	0.744	0.240	0.135	–	–	–	–
Ours	**25.031**	**0.844**	**0.289**	**0.086**	**21.547**	**0.816**	**0.227**	**0.057**	**21.654**	**0.550**	**0.557**	**0.225**

Table 1. **Quantitative comparisons on synthetic and real-world datasets.** Our method consistently achieves the best performance across all metrics and settings, demonstrating superior realism and physical coherence in scene-dynamics simulations.

LPIPS [18], DreamSim [5], PSNR, and SSIM, averaging three generations per sample.

Baselines. We compare with methods that can use static-scene videos as input. *CVX SDEdit* edits the static video with CogVideoX [16] using a text prompt for the target interaction. *CVX-Fun Fine-tuned* fine-tunes the CogVideoX-Fun inpainting model [1] without hand-mesh conditioning, isolating the effect of explicit dexterous action input. For static-camera real videos, we also compare with InterDyn [2], which uses hand-mask guidance.

Quantitative Results. Table 1 shows that DWM achieves the best performance across all evaluated settings. On synthetic dynamic-camera interactions, DWM improves PSNR from 20.541 to 25.031 and DreamSim from 0.175 to 0.086 over the strongest baseline. On real-world static-camera videos, it also outperforms InterDyn and the fine-tuned inpainting baseline, indicating that hand-mesh conditioning models object dynamics rather than merely inserting plausible hands. DWM further improves all metrics on real-world dynamic-camera data, suggesting that aligned synthetic egocentric supervision and fixed-camera real dynamics combine effectively.

Qualitative Results. Figure 3 compares synthetic and real-world scenes under dynamic egocentric views. Text-guided SDEdit struggles to perform the intended contact-rich action, while the fine-tuned inpainting baseline often misses the target object or hallucinates incorrect dynamics. DWM follows the egocentric hand trajectory, manipulates the intended target, and keeps the rest of the scene stable. Notably, it generalizes to completely unseen real-world dynamic-view scenes and produces coherent action-conditioned dynamics, such as opening a sliding window, despite the absence of such interactions in training. This matches the intended digital-twin setting: reconstruct once, then animate the scene with new dexterous actions.

4. Conclusion

We presented Dexterous World Models, a scene-action-conditioned video diffusion framework for simulating egocentric interaction dynamics in static 3D scenes. DWM separates the static environment from manipulation-induced residual changes by conditioning on aligned static-scene renderings and egocentric hand-mesh videos. Initialized from an inpainting prior and trained with hybrid synthetic-real supervision, the model preserves scene consistency while generating plausible dexterous dynamics. Experiments show strong synthetic-to-real generalization and realistic hand-guided object dynamics in static digital twins.

References

[1] aigc-apps. VideoX-Fun, 2024. https://github.com/aigc-apps/VideoX-Fun. 2, 3, 4

[2] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. InterDyn: Controllable interactive dynamics with video diffusion models. In *Proc. CVPR*, 2025. 2, 4

[3] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proc. CVPR*, 2025. 2

[4] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561*, 2023. 3

[5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 4

[6] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018. 1

[7] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proc. ICCV*, 2024. 2, 3

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 3

[9] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Proc. ICLR*, 2022. 2

[10] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proc. ICCV*, 2019. 3

[11] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proc. CVPR*, 2024. 3

[12] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, 2023. 3

[13] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), 2017. 3

[14] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proc. CVPR*, 2025. 2

[15] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 2

[16] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *Proc. ICLR*, 2025. 4

[17] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for Gaussian splatting. *J. Machine Learning Research*, 26(34), 2025. 3

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 2018. 4

[19] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. TASTE-Rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proc. CVPR*, 2025. 2, 3

[20] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proc. ICCV*, 2025. 2