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

Recent progress in 3D reconstruction has made it easy to create realistic digital twins from everyday environments. However, current digital twins remain largely static, limited to navigation and view synthesis without embodied interactivity. To bridge this gap, we introduce Dexterous World Model (DWM), a scene-action-conditioned video diffusion framework that models how dexterous human actions induce dynamic changes in static 3D scenes. Given a static 3D scene rendering and an egocentric hand motion sequence, DWM generates temporally coherent videos depicting plausible human-scene interactions. Our approach conditions video generation on (1) static scene renderings following a specified camera trajectory to ensure spatial consistency, and (2) egocentric hand mesh renderings that encode both geometry and motion cues in the egocentric view to model action-conditioned dynamics directly. To train DWM, we construct a hybrid interaction video dataset: synthetic egocentric interactions provide fully aligned supervision for joint locomotion-manipulation learning, while fixed-camera real-world videos contribute diverse and realistic object dynamics. Experiments demonstrate that DWM enables realistic, physically plausible interactions, such as grasping, opening, or moving objects, while maintaining camera and scene consistency. This framework establishes the first step toward video diffusion-based interactive digital twins, enabling embodied simulation from egocentric actions.

1. Introduction

World models [14] aim to capture the underlying structure and dynamics of the environment so that intelligent systems can reason, plan, and act effectively. Modeling dynamic consequences and causal relationships is central to world modeling, which enables the prediction of environmental changes

[*]Equal contribution

resulting from a wide range of actions. Notably, many of the most important and complex dynamics are driven by human-environment interactions, where humans manipulate objects and actively alter the state of the world. A practical world model must therefore: (1) maintain a consistent representation of the static components of a scene, (2) accept an action specification (e.g., dexterous hand manipulation) that induces dynamic changes, and (3) generate the resulting dynamics while faithfully preserving the unaltered regions.

While video generative models [52] provide a promising foundation for world modeling, they remain insufficient to function as a faithful world model. These models typically attempt to synthesize the entire scene, both static and dynamic regions, making it challenging to focus on the specific dynamic changes induced by action inputs. Furthermore, most video models [3, 42, 52] accept only textual instruction as "action input". However, textual descriptions are inherently imprecise, as they cannot accurately specify human actions such as hand configurations, fine-grained temporal control, both of which are crucial for modeling realistic human manipulations. Many existing world models often treat camera motion as the action input [6, 63]. However, camera motion is not the primary drivers of world dynamics. The most significant source of changes in everyday settings is dexterous hand interaction, such as grasping and object manipulation, which are absent from existing approaches.

In this paper, we present **Dexterous World Models (DWM)**, a new formulation of world modeling that takes both a static scene and dexterous human hand motions as input, and predicts the dynamic changes induced by those actions, as shown in Fig. 1. Unlike existing approaches that require the model to synthesize the entire scene, both static and dynamic, we explicitly provide a rendering of the static environment as input, together with dexterous hand trajectories captured from an egocentric perspective. This design reflects a more natural setting because humans also perceive and internalize the static world before acting upon it, encouraging the model to focus on the dynamic changes of the world induced by human actions, while preserving others. The egocentric formulation further provides inherent grounding for interaction because it naturally captures the user's focus of attention and hand trajectory during manipulation.

To realize our new formulation, we implement DWM as a *scene-action-conditioned video diffusion model* that enables action-conditioned simulation in static 3D scenes. DWM is designed to capture how human actions induce dynamic changes in an otherwise static environment, directly modeling realistic scene interactions driven by dexterous hand actions. Our approach conditions the diffusion process on two inputs: (1) a static scene video rendered along a specified camera trajectory to ensure spatial consistency, and (2) egocentric hand mesh renderings that encode geometric and motion cues for modeling action-conditioned dynamics. A core idea behind our formulation is to model only the residual dynamics caused by human actions while preserving all unaltered regions. To achieve this, we initialize DWM from a pretrained video inpainting diffusion model [1]. When given a full mask, such models reproduce the input video, effectively behaving as an identity mapping with a generative prior. Under this initialization, using the static scene video as input establishes a base output that preserves the egocentric camera motion and scene appearance without introducing interaction dynamics. Conditioning the model on dexterous egocentric hand motions then guides the diffusion process to synthesize only manipulation-induced changes.

As no existing dataset provides paired static scene renderings and corresponding interaction videos captured under identical camera trajectories required by our formulation, we leverage a synthetic 3D human–scene interaction dataset [20] to construct aligned interaction videos, static scene videos, and hand videos. While synthetic data provides precise alignment under identical camera trajectories, it offers limited diversity in interaction dynamics. We therefore additionally incorporate fixed-camera real-world interaction videos [58], enriching the model with diverse and realistic interaction dynamics. Our experiments show that DWM generalizes to real-world scenarios involving combined locomotion and manipulation. Beyond video synthesis, we demonstrate that DWM can be used as a visual world model for simulation-based action evaluation by comparing the predicted visual outcomes of candidate actions against a given goal.

Our main contributions can be summarized as follows: (1) We introduce Dexterous World Models (DWM), a new formulation of world modeling via scene-action-conditioned video diffusion. (2) We propose a conditioning scheme that leverages static scene videos and egocentric hand motions to model interaction-induced residual dynamics. (3) We construct a hybrid training dataset combining synthetic egocentric interactions with fixed-camera real-world interaction videos, which provides aligned supervision and realistic interaction dynamics. (4) We show that DWM can be used for simulation-based reasoning by simulating the visual consequences of candidate actions.

2. Related Work

World Models. World models aim to learn predictive environment dynamics, enabling agents to anticipate how the world evolves in response to actions. The classical formulation of Ha and Schmidhuber [14] demonstrated that compact latent dynamics can support imagination-based planning, inspiring extensive work that integrates learned world models with reinforcement learning for control and decision-making [15, 16, 60]. While effective, these approaches typically operate in low-dimensional latent spaces and do not model rich visual interactions. To capture full visual scene evolution beyond latent dynamics, recent work adopts video generative models to build visual world models that predict future observations directly in pixel space. Action conditioning in these models commonly relies on camera pose changes [6] or Plücker-ray encodings [61, 63], and is often augmented with depth or point-map observations [40, 54, 55] or joint RGB–geometry generation [49, 59, 63] to enforce 3D consistency. More recent work expands the action space to include human body motion [5, 40] or robot action [29, 46]. These approaches commonly adopt an image-to-video paradigm that hallucinates both the scene and its evolution from a single frame, entangling background synthesis with action-driven dynamics. Meanwhile, world models designed for robot-arm manipulation [13, 19, 62] typically assume a fixed camera viewpoint, and therefore cannot model tasks that require both navigation and manipulation. Beyond these approaches, video diffusion models have also been explored as simulators for training embodied agents in driving [18, 38], game environments [4, 9, 41], or text-generated worlds [7]. However, these approaches rely on abstract or limited action spaces and do not model fine-grained, contact-rich human–scene interactions. In contrast, our approach assumes a known static environment and focuses on modeling interaction-induced visual dynamics from dexterous hand actions, while preserving the static scene.

Interactive 3D Scene Reconstruction. Recent advances in 3D reconstruction [44, 45] have made dense geometry and camera pose recovery from images increasingly accessible, lowering the barrier to creating static 3D scenes. However, most existing methods still produce static, monolithic scene representations that lack object-level disentanglement. To enable interaction and editing, recent work explores decompositional neural reconstruction [30, 31, 48, 50] and articulated scene modeling, either by estimating articulation structures [24, 25, 27, 47] or reconstructing full geometry and appearance using 3D Gaussian Splatting [26, 51]. These approaches often require additional supervision such as joint-axis estimates [35, 39] or detailed segmentation [23], limiting scalability and generalization. In contrast, our approach focuses on modeling everyday dexterous human–object interactions directly from static 3D scenes. By conditioning a video diffusion model on egocentric camera motion and fine-grained hand manipulations, we simulate temporally coherent interaction dynamics without explicit articulation modeling, bringing visual interactivity to static digital twins.

3. Method

3.1. Formulation of Dexterous World Models

World models [14] aim to predict how actions transform the surrounding environment, allowing agents to simulate or anticipate future outcomes. A general formulation of a world model can be written as:

$$p_\theta(\mathbf{S}_{1:F} \mid \mathbf{S}_0, \mathcal{A}_{1:F}) = p_\theta\left((\mathbf{S}_0 + \Delta\mathbf{S}_t)_{t=1:F} \mid \mathbf{S}_0, \mathcal{A}_{1:F}\right), \quad (1)$$

where $\mathbf{S}_0$ denotes the initial scene environment and $\mathcal{A}_{1:F}$ represents the sequence actions applied to the world. $\mathbf{S}_{1:F}$ are the resulting environment states after the actions. To emphasize the dynamic component, we express each future state as $\mathbf{S}_t = \mathbf{S}_0 + \Delta\mathbf{S}_t$, where $\Delta\mathbf{S}_t$ captures the action-induced residual changes relative to the initial static scene.

Existing navigation-only visual world models [6, 63] primarily focus on large-scale navigation and novel view prediction, where actions $\mathcal{A}_{1:F}$ mainly correspond to camera motion $\mathcal{C}_{1:F}$, that is $p_\theta(\mathbf{V}_{1:F} \mid \mathbf{I}_0, \mathcal{C}_{1:F})$. The $\mathbf{V}_{1:F}$ denotes future visual observations and $\mathbf{I}_0$ is the initial frame that provides a partial 2D observation of the underlying static scene $\mathbf{S}_0$. These models aim to predict unseen views $\mathbf{V}_t = \Pi(\mathbf{S}_0; \mathcal{C}_t)$, where Π is a rendering function under camera view $\mathcal{C}_t$. This formulation often assumes a static environment with no state change, expressed as $\Delta\mathbf{S} = \varnothing$.

Alternatively, recent work [5, 40] attempts to incorporate human actions into dynamic scene modeling by expanding $\mathcal{A}_{1:F} = \{\mathcal{C}_{1:F}, \mathcal{H}_{1:F}\}$, where $\mathcal{H}$ denotes human actions. However, this formulation remains suboptimal for modeling dynamics because the model must synthesize both scene appearance and dynamic changes as an entangled representation. More formally, such models can be written as $p_\theta\left(\mathbf{V}_{1:F} \mid \mathbf{I}_0, \mathcal{A}_{1:F} = \{\mathcal{C}_{1:F}, \mathcal{H}_{1:F}\}\right)$, and it can be factorized by marginalizing over the latent static scene $\mathbf{S}_0$ and action-induced dynamics $\Delta\mathbf{S}_{1:F}$:

$$\begin{aligned} p_\theta(\mathbf{V}_{1:F} \mid \mathbf{I}_0, \mathcal{A}_{1:F}) = \int_{\mathbf{S}_0, \Delta\mathbf{S}} & p^d_\theta(\mathbf{S}_0, \Delta\mathbf{S}_{1:F} \mid \mathbf{I}_0, \mathcal{A}_{1:F}) \\ & \cdot p^o_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \Delta\mathbf{S}_{1:F}, \mathbf{I}_0, \mathcal{A}_{1:F})\, d\mathbf{S}_0\, d\Delta\mathbf{S}_{1:F}. \end{aligned} \quad (2)$$

Here, p^d_θ is the *dynamics model* that samples the scene and action-induced changes $(\mathbf{S}_0, \Delta\mathbf{S}_{1:F})$ given the actions $\mathcal{A}_{1:F} = \{\mathcal{C}, \mathcal{H}\}_{1:F}$, while p^o_θ is the *observation model* that maps latent world states to visual frames $\mathbf{V}_{1:F}$. The observation model can be simplified to $p^o_\theta\left(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \Delta\mathbf{S}_{1:F}, \mathcal{C}_{1:F}\right)$, as the visual outcome $\mathbf{V}_{1:F}$ is conditionally independent from the action input $\{\mathcal{H}\}_{1:F}$ and image observation $\mathbf{I}_0$, given the known dynamics $\Delta\mathbf{S}_{1:F}$ and entire static scene $\mathbf{S}_0$. Similarly, in the dynamics model, we may drop the camera motion since it does not affect the world dynamics: $p^d_\theta\left(\mathbf{S}_0, \Delta\mathbf{S}_{1:F} \mid \mathbf{I}_0, \mathcal{H}_{1:F}\right)$. However, since this dynamics model is still required to generate $\mathcal{S}_0$ as part of the output conditioned on the action sequence, it becomes vulnerable to altering the static world instead of focusing on the desired dynamic state changes $\Delta\mathbf{S}_{1:F}$. This coupling of scene generation and dynamics breaks causal consistency, making accurate dynamic modeling difficult.

In our DWM formulation, we explicitly separate the static world $\mathcal{S}_0$ from its dynamics by conditioning the model on $\mathcal{S}_0$. Our dexterous world model thus learns:

$$p_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \mathcal{A}_{1:F}) = p_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \{\mathcal{C}_{1:F}, \mathcal{H}_{1:F}\}), \quad (3)$$

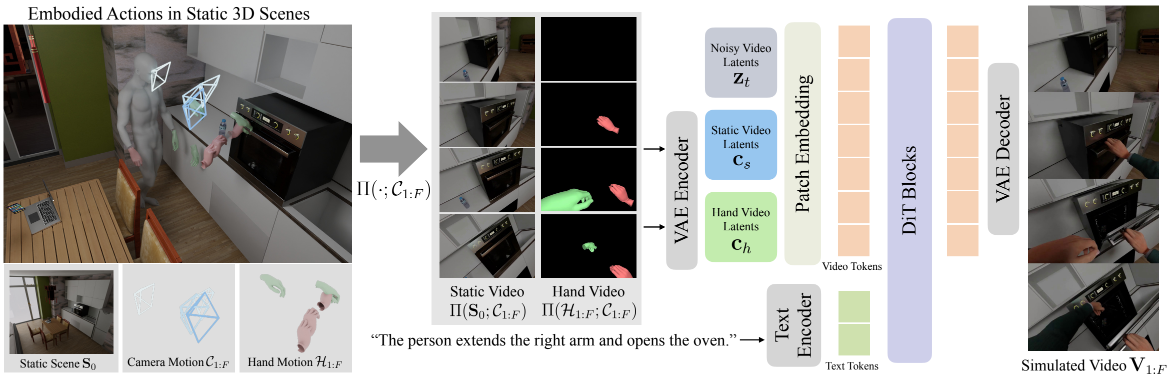


Figure 2. **Overview.** DWM simulates egocentric visual dynamics induced by embodied actions within a given static 3D scene. We instantiate it as a video diffusion model conditioned on the egocentric projections of the static scene and hand trajectories.

where $\mathbf{S}_0$ remains fixed and only its state evolves according to the applied actions. By introducing temporal scene dynamics $\Delta\mathbf{S}_{1:F}$ as a latent variable and marginalizing over it, this can be expressed as

$$\begin{aligned} p_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \mathcal{A}_{1:F}) &= \int_{\Delta\mathbf{S}} p^d_\theta(\Delta\mathbf{S}_{1:F} \mid \mathbf{S}_0, \mathcal{A}_{1:F})\, p^o_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \Delta\mathbf{S}_{1:F}, \mathcal{A}_{1:F})\, d\Delta\mathbf{S} \\ &= \int_{\Delta\mathbf{S}} p^d_\theta(\Delta\mathbf{S}_{1:F} \mid \mathbf{S}_0, \mathcal{H}_{1:F})\, p^o_\theta(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \Delta\mathbf{S}_{1:F}, \mathcal{C}_{1:F})\, d\Delta\mathbf{S}. \end{aligned} \tag{4}$$

Unlike the previous modeling in Eq. (2), our dynamics model p^d_θ only produces the dynamic changes $\Delta\mathbf{S}_{1:F}$ conditional to the action inputs $\mathcal{H}_{1:F}$. The observation model p^o_θ then translates this evolving world, described by $(\mathbf{S}_0, \Delta\mathbf{S}_{1:F})$, into a sequence of visual frames observed along the camera trajectory $\mathcal{C}_{1:F}$, which is identical to the observation model in Eq. (2). This structure defines a clear causal process: manipulation drives state transitions in the world, and the camera trajectory determines how these changes appear visually.

3.2. Scene-Action-Conditioned Video Diffusion

Building on the formulation in Eq. (3), we instantiate our framework as an *egocentric, scene-action-conditioned video diffusion model*. Given a static 3D scene $\mathbf{S}_0$ and a sequence of embodied actions $\mathcal{A}_{1:F}$, the model generates a temporally coherent video that visualizes plausible human-scene interactions with dynamic scene changes. We additionally condition the model on a text prompt $\mathcal{T}$, which supplies semantic guidance to the video diffusion process. To ground this generative process in embodied perception, we decompose $\mathcal{A}_{1:F} = \{\mathcal{C}_{1:F}, \mathcal{H}_{1:F}\}$ into a camera trajectory $\mathcal{C}_{1:F}$ and a hand manipulation trajectory $\mathcal{H}_{1:F}$, and assume a conditional independence structure such that the visual sequence $\mathcal{V}_{1:F}$ depends on the static world $\mathbf{S}_0$ and actions $\mathcal{A}_{1:F}$ only through their egocentric renderings:

$$p(\mathbf{V}_{1:F} \mid \mathbf{S}_0, \mathcal{A}_{1:F}, \mathcal{T}) \approx p(\mathbf{V}_{1:F} \mid \Pi(\mathbf{S}_0; \mathcal{C}_{1:F}), \Pi(\mathcal{H}_{1:F}; \mathcal{C}_{1:F}), \mathcal{T}), \tag{5}$$

where $\Pi(\cdot; \mathcal{C}_{1:F})$ denotes the 2D views observed along the egocentric camera trajectory $\mathcal{C}_{1:F}$. This approximation grounds the dynamics in embodied perception, enforcing a causal relation where manipulation modifies the visual state of the scene, while the camera trajectory determines how those changes are observed. Importantly, we instantiate the generative process as a latent video diffusion model [36] conditioned on two egocentric signals: (1) a static-scene video, $\Pi(\mathbf{S}_0; \mathcal{C}_{1:F})$, that enforces spatial consistency along the specified camera trajectory, and (2) a egocentric hand-mesh rendering video [32, 37], $\Pi(\mathcal{H}_{1:F}; \mathcal{C}_{1:F})$), which encodes the manipulation actions that drive the dynamic scene changes. The model then simulates the resulting environment states induced by the hand action inputs under the same camera views, $\mathbf{V}_{1:F} = \Pi(\mathbf{S}_0 + \Delta\mathbf{S}_{1:F}; \mathcal{C}_{1:F})$). See Fig. 2 for the overview of our framework.

Inpainting Priors for Residual Dynamics Learning. We formulate the generation task as learning the residual visual change induced by hand manipulation action, rather than reconstructing the entire video from scratch. Given the target distribution $p(\mathbf{V}_{1:F} \mid \Pi(\mathbf{S}_0; \mathcal{C}_{1:F}), \Pi(\mathcal{H}_{1:F}; \mathcal{C}_{1:F}))$, we decompose it as:

$$\mathbf{V}_{1:F} = \Pi(\mathbf{S}_0 + \Delta\mathbf{S}_{1:F}; \mathcal{C}_{1:F})) = \Pi(\mathbf{S}_0; \mathcal{C}_{1:F}) + \Delta\mathbf{V}_{1:F}, \tag{6}$$

where $\Delta\mathbf{V}_{1:F}$ denotes the action-induced residual dynamics in 2D rendering space. Notably, when $\Delta\mathbf{S}_t$ affects only a small region or induces marginal changes, the corresponding $\Delta\mathbf{V}_t$ is similarly small, and the resulting frames remain close to the static rendering, $\Pi(\mathbf{S}_0; \mathcal{C}_{1:F})$. Thus, an effective model for this task should combine an *identity mapping* capability for preserving static regions with strong *generative priors* for modeling dynamically changing areas.

Motivated by this observation, we reinterpret a pretrained video inpainting diffusion model [1] as *an identity function with generative prior* and adopt it as the initialization of our dexterous world model. Inpainting models are trained to reconstruct masked regions conditioned on visible context. When the mask is set to $m = 1$ (all pixels known), the model behaves as a near-identity operator that faithfully reproduces the input video. Crucially, this is not a trivial identity: the pretrained inpainting network ϵ_{θ_0} has learned to restore spatial structure, temporal smoothness, and appearance continuity, giving it a strong generative prior even under fully observed inputs. By initializing DWM from such an inpainting model, we encourage the diffusion process to preserve static scene appearance and camera motion, while learning to generate only the interaction-induced residual dynamics $\Delta\mathbf{V}_{1:F}$. This formulation leverages the identity-preserving behavior of the inpainting model to stabilize training while learning interaction-induced residual dynamics.

Training. The model operates in the latent space of a pretrained video variational autoencoder (VAE) [22, 52], which encodes an input video into latent tensors $\mathbf{z}_0 \in \mathbb{R}^{f\times c\times h\times w}$ following the encoder's spatiotemporal compression ratio. At each diffusion timestep t, the Diffusion Transformer (DiT) [34] ϵ_θ predicts the noise component from the noisy latent $\mathbf{z}_t$ conditioned on the static scene video latents $\mathbf{c}_s$ and hand mesh video latents $\mathbf{c}_h$. The training objective follows the standard latent diffusion loss [36]:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{z}_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t \mid \mathbf{c}_s, \mathbf{c}_h)\|_2^2\right]. \tag{7}$$

The conditional latents are concatenated with $\mathbf{z}_t$ along the channel dimension before being processed by ϵ_θ. During inference, the model iteratively denoises $\mathbf{z}_t$ to obtain $\hat{\mathbf{z}}_0$, which is decoded by VAE into a realistic interaction video.

3.3. Paired Interaction Video Dataset Construction

Training our model requires paired data consisting of (i) a static scene video rendered along a camera trajectory, (ii) an egocentric hand-mesh video, and (iii) a corresponding interaction video depicting action-induced scene changes. Collecting such triplets in real-world environments is challenging in practice, as it would require capturing both a static scene and a dynamic human–object interaction under a perfectly aligned egocentric camera trajectory. This alignment is essential for learning how human actions transform a static scene while preserving spatial consistency. Due to the practical difficulty of obtaining such paired data under dynamic egocentric viewpoints in the real world, we construct a hybrid training dataset. Specifically, we combine synthetic egocentric interaction video pairs, which provide precise alignment, with fixed-camera real-world interaction videos, which offer realistic physical dynamics. In addition, to evaluate generalization to real-world settings with dynamic viewpoints, we develop a dedicated data construction protocol to collect a real-world dataset for evaluation.

Synthetic Interaction-Static Scene Video Pairs. We leverage a synthetic 3D human-scene interaction dataset, TRUMANS [20], which offers full control over human actions, scene states, and camera trajectories, allowing exact alignment between static and interactive sequences. The actor is parameterized using SMPL-X [32], which provides full-body pose control. We place a virtual camera at the midpoint between the eyes in the canonical pose, aligned with the head's forward axis and rigidly attached to the head joint, to simulate egocentric viewpoints. For each prerecorded action sequence in TRUMANS, we render three synchronized outputs: (1) the interaction video captures the full human–scene interaction from the moving egocentric camera; (2) the static-scene video is rendered by replaying the same camera trajectory over the static environment without applying any object state changes; and (3) the hand-mesh video is obtained by rendering only the hand surfaces, segmented from the full-body mesh, along the same trajectory. This setup ensures spatiotemporal alignment across all videos, providing a precise supervisory signal for learning manipulation-induced visual dynamics.

Fixed-Camera Real-World Interaction Videos. While synthetic data provides precise alignment, it lacks the complex physical effects observed in the real world, such as fluid dynamics, material deformation, and subtle visual variations. To incorporate such effects, we additionally use real-world human–object interaction videos recorded from fixed cameras [58], where $\mathcal{C}_t = \mathcal{C}_0$ for all $t \in 1,\ldots,F$. Under this setup, the egocentric static scene video is time-invariant:

$$\Pi(\mathbf{S}_0;\mathcal{C}_t) = \Pi(\mathbf{S}_0;\mathcal{C}_0) = \mathbf{V}_0, \quad \forall t \in \{1,\ldots,F\}. \tag{8}$$

Therefore, by simply repeating the first frame $\mathbf{V}_0$ across F frames, we construct a static-scene video that aligns exactly with the interaction video in viewpoint. Combined with predicted hand-mesh renderings $\Pi(\mathcal{H}_{1:F};\mathcal{C}_{1:F})$, obtained via HaMeR [33], this enables us to construct training triplets $(\mathbf{V}_{1:F}, \Pi(\mathbf{S}_0;\mathcal{C}_{1:F}), \Pi(\mathcal{H}_{1:F};\mathcal{C}_{1:F}))$. Notably, existing egocentric datasets [8, 11, 17, 58] do not explicitly provide such pairs aligned under shared camera motion. Our construction leverages fixed-camera setups to approximate this pairing, enabling supervision without requiring separate static captures or 3D scene reconstruction. This extends dexterous world modeling to real-world data while maintaining compatibility with our observation model in Eq. (5).

Real-World Captures for Evaluation. Although real-world paired data under dynamic egocentric viewpoints is impractical to obtain at scale, evaluating generalization to such settings is essential for validating a world model. To this end, we develop a dedicated data construction protocol to collect a real-world dataset for evaluation under dynamic viewpoints. We use Aria [10], which provides millimeter-level camera trajectory estimates via built-in SLAM. During capture, the operator first explores the scene before performing the interaction, enabling reconstruction of a 3D Gaussian representation of the static scene from pre-action frames [21, 53]. The reconstructed scene is then rendered along the recorded camera trajectory during interaction, yielding paired static-scene videos aligned with the corresponding ground-truth interaction sequences. Using this protocol, we collect 60 paired real-world samples under dynamic egocentric viewpoints, covering a wide range of interaction types, including simple and complex pick-and-place, articulated object manipulation (e.g., opening a washing machine, folding a chair), and counterfactual dynamics (e.g., pressing an elevator button and the door opens, turning on a faucet and water flows). Examples are shown in our Supp. Mat..

3.4. Action Evaluation with DWM

Given a static 3D scene $\mathbf{S}_0$ and a set of action candidates $\{\mathcal{A}^{(i)}\}_{i=1}^{N}$, we simulate their visual consequences using DWM as described in Eq. (3), producing videos $\{\mathbf{V}_{1:F}^{(i)}\}_{i=1}^{N}$. We then rank the actions according to how well their predicted visual outcomes align with a given goal, specified either as a language instruction or an image. For language-based goals, we compute the semantic similarity between each generated video $\mathbf{V}_{1:F}^{(i)}$ and the goal text g_{text} using a pretrained VideoCLIP model [43]:

$$s_{\text{text}}^{(i)} = \text{sim}_{\text{VC}}(\mathbf{V}_{1:F}^{(i)}, g_{\text{text}}), \tag{9}$$

where $\text{sim}_{\text{VC}}(\cdot, \cdot)$ denotes the cosine similarity between video and text embeddings. For image-based goals, we compare the final frame $\mathbf{I}_F^{(i)} = \mathbf{V}_F^{(i)}$ of each simulation against a goal image $\mathbf{I}_{\text{goal}}$ using the LPIPS perceptual metric [57]:

$$s_{\text{img}}^{(i)} = -\text{LPIPS}(\mathbf{I}_F^{(i)}, \mathbf{I}_{\text{goal}}), \tag{10}$$

where lower LPIPS indicates higher perceptual similarity. The best action is selected as the one maximizing the score:

$$\mathcal{A}^* = \arg\max_{\mathcal{A}^{(i)}} s_{\text{text}}^{(i)} \quad \text{or} \quad \mathcal{A}^* = \arg\max_{\mathcal{A}^{(i)}} s_{\text{img}}^{(i)}, \tag{11}$$

depending on the type of the given goal. This formulation enables goal-driven action selection via simulation, without requiring explicit reward functions or real-world trials.

4. Experiments

Dataset. We construct a benchmark of 144 samples, each consisting of a static scene video, a hand-mesh video, a ground-truth interaction video. To ensure comprehensive evaluations, the benchmark includes both synthetic and real-world data. For the synthetic subset (Synthetic Dynamic), we sample 48 sequences from the TRUMANS [20] that are excluded from training. The real-world subset includes 96 videos under static or dynamic camera settings. In the static camera setting (Real-World Static), we use 48 samples unseen during training from TASTE-Rob [58]. For the dynamic camera setting (Real-World Dynamic), we use 48 samples from our custom captured dataset by Aria Glasses.

Baselines and Metrics. We compare our approach against two baselines that can directly utilize static scene videos as conditioning inputs. First, We apply SDEdit [28] to static videos using text-prompt guidance describing the target action and resulting scene changes (*CVX SDEdit*). SDEdit first adds noise to the input video, then denoises it through the SDE prior under text conditioning for video editing. We use CogVideoX [52] for denoising and set the noise strength to 0.75 with 50 inference steps. Second, we evaluate a fine-tune version of our base model, CogVideoX-Fun [1] (*CVX-Fun Fine-tuned*). CogVideoX-Fun is an inpainting-based video diffusion model that takes an input video, a binary mask sequence specifying the inpainting region, and a text prompt to generate inpainted videos. We set the mask as ones and fine-tune the model with our dataset without the hand-mesh video condition. For the static camera setting, we additionally compare against InterDyn [2], which synthesizes hand–object interaction videos from an initial frame using a hand-mask video injected into the diffusion process via ControlNet [56].

We evaluate generated videos using both perceptual and pixel-level metrics against ground-truth videos. Specifically, we report LPIPS [57] and DreamSim [12] scores for perceptual similarity, and PSNR and SSIM for pixel-level quality. For each sample in our benchmark, we generate three videos with random seeds and report the averaged results.

4.1. Comparison

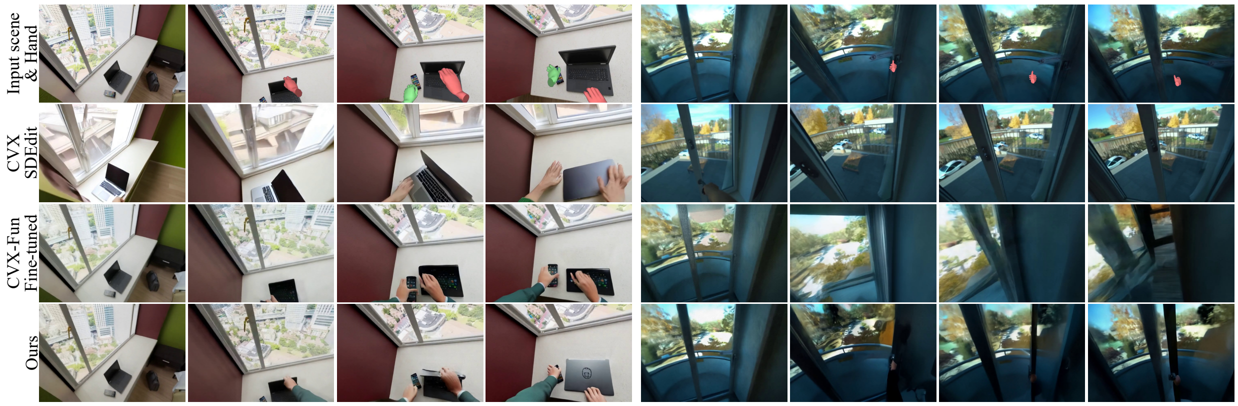
Tab. 1 shows quantitative performance across synthetic and real-world datasets under static or dynamic cameras, complementing the qualitative results presented in Figs. 3 and 4.

Real-World Scene with Static Camera. Fig. 4 compares our method with baselines in a real-world scene with a static camera. CVX-SDEdit struggles to generate meaningful interactions while maintaining scene appearance. CVX-Fun Fine-tuned often fails to interact with the correct target and hallucinates the object of interest. InterDyn [2] produces spatially aligned hands with the input masks but fails to model resulting object dynamics. In contrast, our model synthesizes physically plausible interactions in which hand actions induce consistent scene changes, such as object displacement or articulation. Quantitatively, ours outperforms baselines across all metrics on real-world static camera videos.

Scenes with Dynamic Camera. Fig. 3 shows comparisons on synthetic and real-world scenes captured in dynamic views. Despite real-world scenes with dynamic view being completely unseen during training, and the absence of any training samples involving opening a window, DWM successfully generates coherent simulations, demonstrating strong generalization. This strong generalization is quantitatively validated in Tab. 1, where our method achieves superior performance across all metrics.

	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.



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

Figure 3. **Qualitative comparison on synthetic and real-world scenes with dynamic view.** DWM successfully generates physically plausible simulations with dynamic view changes corresponding to the input hand actions. Notably, our method generalizes well to completely unseen real-world scenes, producing coherent action-conditioned dynamics such as opening a sliding window.

Figure 4. **Qualitative comparison on real-world scenes with static camera.** Our method produces realistic interactions with consistent scene dynamics. Baselines fail to perform meaningful actions or hallucinate incorrect interactions.

4.2. Qualitative Results

Navigation-Manipulation Disentanglement. Fig. 5 demonstrates DWM's ability to disentangle navigation and manipulation. Without hand motion conditioning, DWM operates as a pure navigator. When conditioned on hand motion, it additionally models manipulation-induced dynamics, generating realistic scene changes driven by hand actions.

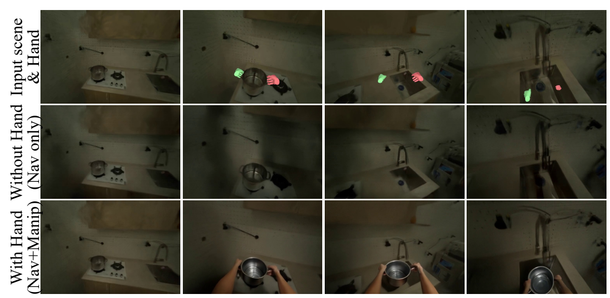


Figure 5. **Navigation-manipulation disentanglement.** Without hand-motion input, DWM simulates navigation only. Conditioning on hand motion enables the model to generate action-induced visual dynamics, highlighting navigation-manipulation disentanglement.

Accurate Targeting. Fig. 6 demonstrates DWM's ability to precisely select and manipulate the target according to different hand-action conditions. The model responds to variations in hand motion, resulting in accurate interactions.

Action Evaluation. Fig. 7 illustrates how DWM can be used as an action evaluator. The model simulates outcomes for multiple action candidates and scores each based on its proximity to the desired goal. Depending on the goal specification, we measure either semantic alignment with VideoCLIP [43] (for text goals) or perceptual similarity using LPIPS [57] (for image goals). The action with the highest CLIP similarity or lowest LPIPS distance is selected.

	Synthetic				Real-World							
	Dynamic Camera				Static Camera				Dynamic Camera			
	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
TRUMANS Only	24.151	0.834	0.304	0.093	17.959	0.766	0.304	0.124	20.654	0.520	**0.543**	0.273
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 2. **Ablation on training data composition.** Including real-world data for training significantly improves both perceptual and pixel-level metrics across synthetic and real-world test sets, demonstrating that static-camera real-world interactions enhance generalization.

	PSNR↑	SSIM↑	LPIPS↓	DreamSim.↓
AdaLN (global)	21.962	0.806	0.306	0.127
AdaLN (per frame)	22.789	0.827	**0.288**	0.110
Hand Mask	22.876	0.797	0.338	0.137
Ours	**24.151**	**0.834**	0.304	**0.094**

Table 3. **Ablation on hand-motion conditioning.** We demonstrate the effectiveness of spatially aligned hand-mesh conditioning.

Figure 6. **Accurate targeting.** DWM accurately manipulates the intended target object based on variations in hand-action condition. Input scene and hand conditions are shown in the top-left corners.

Figure 7. **Action evaluation.** DWM simulates multiple candidate actions and ranks them based on goal alignment.

4.3. Ablation Study

Impact of Hybrid Dataset. We evaluate the impact of hybrid training data composition by comparing models trained with only TRUMANS versus those trained with both TRUMANS and TASTE-Rob. As shown in Tab. 2, incorporating real-world data consistently improves performance on most metrics, not only on synthetic test sequences but also on Aria-captured real-world data, which includes both locomotion and fine-grained manipulation under dynamic camera motion. This is notable given that the real-world training data includes only static-camera interactions, demonstrating strong generalization to more complex real-world scenarios. The results validate our training data combination strategy: synthetic data offers controllable paired supervision with diverse camera motion, while real-world videos provide high-fidelity physical dynamics and visual realism.

Hand Motion Conditioning. We compare three types of hand-motion conditioning: parameter-based modulation(AdaLN [34]), hand mask conditioning, and our hand-mesh-rendered conditioning (Tab. 3). AdaLN-based conditioning directly injects MANO [37] pose parameters with global orientation and translation of both hands into the diffusion transformer through Adaptive Layer Normalization. We explore two variants: *AdaLN (Global)*, which aggregates pose features over the entire sequence to provide a single global embedding, and *AdaLN (Per-frame)*, which maintains frame-level embeddings to capture temporal variations. While both allow structured control, their parameter-space conditioning lacks pixel alignment, limiting the model's ability to reproduce detailed hand–scene contact and appearance changes. *Hand mask* instead provides pixel-level alignment by encoding binary hand segmentation videos via the same VAE encoder. However, the coarse silhouettes convey only location cues and lack precise articulation, leading to less coherent motion. *Ours* uses rendered hand-mesh [32, 37] videos as pixel-aligned conditioning signals. Unlike segmentation masks, the rendered meshes provide explicit articulation and contact geometry, allowing the model to directly associate hand motion with appearance changes in the scene. This explicit and spatially consistent conditioning leads to clear improvements, as shown in Tab. 3.

5. Conclusion

We introduced **Dexterous World Models (DWM)**, a video diffusion framework for modeling egocentric visual dynamics induced by locomotion and dexterous hand interactions in a known static 3D scene. Built on an egocentric observation model, DWM predicts residual visual changes conditioned on rendered static scene views and hand mesh trajectories. To train this model, we proposed a hybrid data construction strategy combining synthetic sequences with fixed-camera real-world videos, enabling aligned supervision and physically plausible dynamics. Experiments show that DWM generates realistic and grounded outcomes, generalizes to unseen real-world sequences involving both navigation and manipulation, and effectively disentangles viewpoint shifts from manipulation-induced state changes.

Acknowledgments

This work was supported by RLWRLD, NRF grant funded by the Korean government (MSIT) [No. RS-2022-NR070498, No. RS-2025-25396144, and No. PJT-25-122310], and IITP grant funded by the Korea government (MSIT) [No. RS-2024-00439854, No. RS-2025-25442338, No. RS-2021-II211343, and No. RS-2025-02653113]. H. Joo is the corresponding author.

References

[1] aigc apps, 2024. https://https://github.com/aigc-apps/VideoX-Fun. 2, 4, 6

[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. 6

[3] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 2

[4] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *NeurIPS*, 2024. 3

[5] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. *arXiv preprint arXiv:2506.21552*, 2025. 3

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

[7] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proc. ICML*, 2024. 3

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proc. ECCV*, 2018. 5

[9] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. 3

[10] 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. 5

[11] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proc. CVPR*, 2025. 5

[12] 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. *NeurIPS*, 2023. 6

[13] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *NeurIPS*, 2024. 3

[14] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *NeurIPS*, 2018. 1, 2, 3

[15] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proc. ICML*, 2019. 2

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025. 2

[17] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 5

[18] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 3

[19] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705v2*, 2025. 3

[20] 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, 5, 6

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 5

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014. 5

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. ICCV*, 2023. 3

[24] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proc. ICCV*, 2023. 3

[25] Jiayi Liu, Denys Iliash, Angel X Chang, Manolis Savva, and Ali Mahdavi Amiri. Singapo: Single image controlled generation of articulated parts in objects. In *Proc. ICLR*, 2025. 3

[26] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Artgs: Building interactable replicas of complex articulated objects via gaussian splatting. *arXiv preprint arXiv:2502.19459*, 2025. 3

[27] Zhao Mandi, Yijia Weng, Dominik Bauer, and Shuran Song. Real2code: Reconstruct articulated objects via code generation. In *Proc. ICLR*, 2025. 3

[28] 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. *Proc. ICLR*, 2022. 6

[29] Riccardo Mereu, Aidan Scannell, Yuxin Hou, Yi Zhao, Aditya Jitta, Antonio Dominguez, Luigi Acerbi, Amos Storkey, and Paul Chang. Generative world modelling for humanoids: 1x world model challenge technical report. *arXiv preprint arXiv:2510.07092*, 2025. 3

[30] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. In *NeurIPS*, 2024. 3

[31] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *Proc. CVPR*, 2025. 3

[32] 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. 4, 5, 8

[33] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proc. CVPR*, 2024. 5

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, 2023. 5, 8

[35] Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proc. ICCV*, 2023. 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 4, 5

[37] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):1–17, 2017. 4, 8

[38] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. 3

[39] Xiaohao Sun, Hanxiao Jiang, Manolis Savva, and Angel Chang. Opdmulti: Openable part detection for multiple objects. In *Proc. Intl. Conf. on 3D Vision (3DV)*, 2024. 3

[40] Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. *arXiv preprint arXiv:2506.09995*, 2025. 3

[41] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *Proc. ICLR*, 2025. 3

[42] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[43] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. Videoclip-xl: Advancing long description understanding for video clip models. *arXiv preprint arXiv:2410.00741*, 2024. 6, 7

[44] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proc. CVPR*, 2025. 3

[45] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proc. CVPR*, 2024. 3

[46] Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *Proc. ICCV*, 2025. 3

[47] Yijia Weng, Bowen Wen, Jonathan Tremblay, Valts Blukis, Dieter Fox, Leonidas Guibas, and Stan Birchfield. Neural implicit representation for building digital twins of unknown articulated objects. In *Proc. CVPR*, 2024. 3

[48] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *Proc. ECCV*, 2022. 3

[49] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025. 3

[50] Hongchi Xia, Chih-Hao Lin, Hao-Yu Hsu, Quentin Leboutet, Katelyn Gao, Michael Paulitsch, Benjamin Ummenhofer, and Shenlong Wang. Holoscene: Simulation-ready interactive 3d worlds from a single video. *arXiv preprint arXiv:2510.05560*, 2025. 3

[51] Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, and Wei-Chiu Ma. Drawer: Digital reconstruction and articulation with environment realism. In *Proc. CVPR*, 2025. 3

[52] 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. 2, 5, 6

[53] 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. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 5

[54] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proc. ICCV*, 2025. 3

[55] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. ICCV*, 2023. 6

[57] 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. 6, 7

[58] 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, 5, 6

[59] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025. 3

[60] 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. 2

[61] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proc. ICCV*, 2025. 3

[62] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025. 3

[63] 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, 3