# Generated Reality: Human-centric World Simulation using Interactive Video Generation with Hand and Camera Control

Linxi Xie[1,2*†] Lisong C. Sun[1*] Ashley Neall[1,3*†] Tong Wu[1] Shengqu Cai[1] Gordon Wetzstein[1]
[1]Stanford University [2]NYU Shanghai [3]UNC Chapel Hill

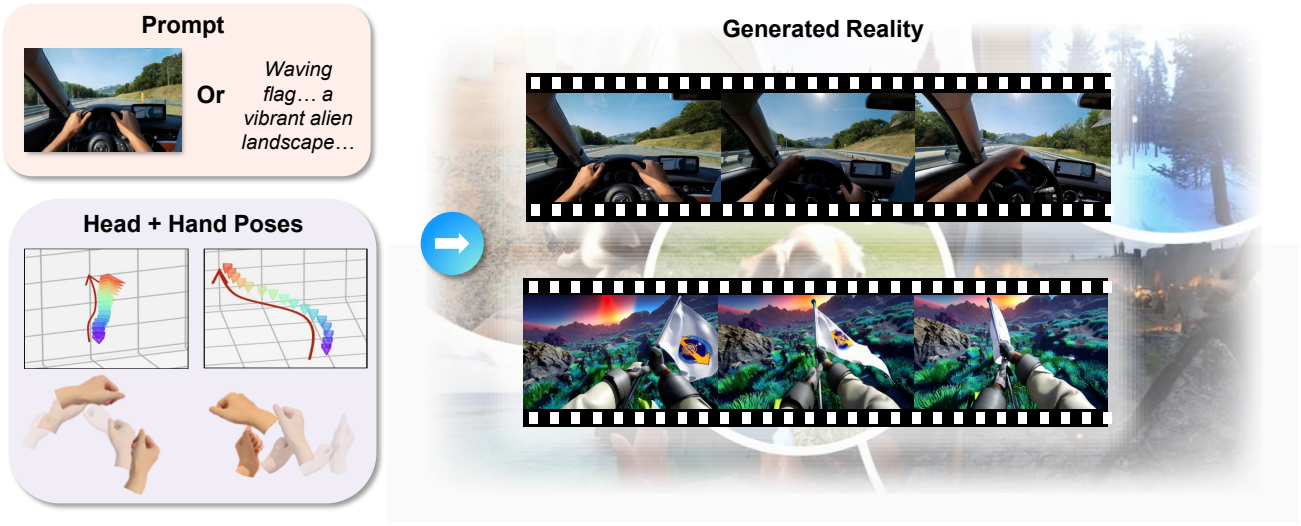



Figure 1. Generated reality is a concept that incorporates human-tracked data (left) into an autoregressive video generation model to enable immersive experiences (right). These generated virtual environments do not rely on laboriously designed 3D assets but are created in a zero-shot manner by the video generator. We explore diffusion transformer conditioning strategies for joint-level hand and head poses, identifying a hybrid 2D–3D strategy as the most effective approach. Our bidirectional attention-based video generator is distilled into a few-step autoregressive model, enabling interactive, human-centric experiences supporting dexterous hand–object interactions.

## Abstract

*Extended reality (XR) demands generative models that respond to users' tracked real-world motion, yet current video world models accept only coarse control signals such as text or keyboard input, limiting their utility for embodied interaction. We introduce a human-centric video world model that is conditioned on both tracked head pose and joint-level hand poses. For this purpose, we evaluate existing diffusion transformer conditioning strategies and propose an effective mechanism for 3D head and hand control, enabling dexterous hand–object interactions. We train a bidirectional video diffusion model teacher using this strategy and distill it into a causal, interactive system that generates egocentric virtual environments. We evaluate this generated reality system with human subjects and demonstrate improved task performance as well as a significantly higher level of perceived amount of control over the performed actions compared with relevant baselines. The project website is at* https://codeysun.github.io/generated-reality/.

## 1. Introduction

Extended reality (XR)—encompassing virtual, augmented, and mixed reality—is crucial in healthcare and rehabilitation, education and professional training, design and engineering, as well as entertainment and media. Despite its transformative potential across these domains, the creation of XR content remains difficult, laborious, and expensive due to the need for specialized expertise, complex development tools, and high production costs.

Emerging video world models offer a powerful platform to address the challenge of content creation for immersive technologies. These large generative AI models are able

*Equal Contribution.
†Work done as a visiting researcher at Stanford.

Wearing **astronaut gloves**, grips the shaft of **a waving flag**... a vibrant **alien landscape** under a colorful sky...

A bright **outdoor park** on a clear day... a friendly **golden retriever** sits obediently...

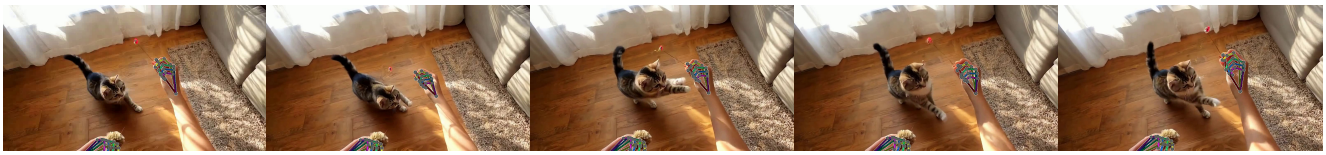
A gritty **medieval dungeon**... the right hand wields a **steel longsword**... an **armored soldier** charges towards the viewer...

A quaint **A-frame cottage**... surrounded by turquoise waters and sandy beaches... palm trees sway in the background...

A lush green **golf course** on a sunny day... hands are **swinging a golf club**.... a golf buggy and a caddy stand ready...

**Pushes the wooden door** open... revealing a magical **winter forest**.. a vintage lamppost glows warmly...

**Driving on a highway** in a modern car... a green countryside with trees on a bright clear summer day...



**Holding a cat wand toy** with a fuzzy pom-pom ball... a **playful domestic cat** swipes repeatedly at the fuzzball...

Figure 2. **Diverse generations.** Leveraging the implicit world knowledge of foundation video models, our system generalizes to diverse scenarios with complex interactions. Generated videos (top) are visualized with input hand conditioning overlaid. Note that, consistent with the pretraining data, input text prompts (below) are augmented with an LLM before being input into the model.

to autoregressively generate close-to-photorealistic video at interactive framerates conditioned on actions or other signals [8, 13, 23, 26, 34].

Current video world models, however, remain limited in the types of conditioning signals they accept, often restricted to simple keyboard controls or text prompts [7, 13, 23, 26, 37, 38, 41]. The limited control makes current world models ineffective as human-centric content generation tools for XR applications. Recent works have focused on conditioning on camera motion [2, 3, 17] or full-body pose [4, 33], showing promise in modeling interactive egocentric dynamics. However, these approaches lack the precision required to represent the detailed wrist and finger movements involved in dexterous hand–object interactions. As a result, it remains an open question how to effectively incorporate joint-level hand pose conditioning into video diffusion models. Furthermore, it is unclear which conditioning strategies best preserve hand fidelity, realism, and temporal coherence in video generation.

We hypothesize that next-generation world models could support truly embodied interactivity by effectively incorporating rich streams of tracked user data, including head and gaze direction, body pose, foot placement, hand and finger articulation, and full-body movement. To this end, we develop a human-centric video world model that enables interactive content generation across both existing and yet-unimagined applications, with a focus on effective head and hand control (see Figure 2). Specifically, we present the first systematic study of hand pose conditioning strategies in video diffusion models. We compare several representative approaches, including token concatenation, addition, cross-attention, ControlNet-style conditioning, and adaptive layer normalization, using metrics that evaluate visual quality and hand-pose fidelity. We find that a combination of 2D ControlNet-style conditioning and a 3D joint-level representation of hand poses injected via token addition is the most effective. Finally, we distill our head- and hand-conditioned video generation model into a causal, real-time architecture, achieving 11 frames per second with a latency of 1.4 seconds on a remotely streamed H100. We conduct a user study with this system, demonstrating significantly improved task performance on three different tasks and a substantially larger perceived sense of control by human subjects compared to relevant baselines.

Our vision of generated reality could enable immersive learning, training, and exploration by allowing users to acquire skills, practice complex tasks without detailed models, and experience real or imagined environments in a zero-shot manner. It could support novel interactive media and real-time generative guidance through smart eyewear for diverse applications.

Our key technical contributions include:

- We conduct a **comprehensive ablation** study comparing hand pose conditioning strategies for video diffusion models, identifying a combination of 2D ControlNet-style conditioning and 3D joint conditioning as the most effective strategy. **Our method outperforms baselines** on video quality, camera pose accuracy, and hand pose accuracy metrics.
- We distill our camera- and hand-conditioned bidirectional teacher model into an **interactive, autoregressive student model** that runs at interactive frame rates. Using this model, we demonstrate **improved task accuracy** and **increased perceived control** in our user studies.

## 2. Related Work

### 2.1. From Video Generation to World Simulation

Recent progress in diffusion models has significantly advanced the field of video generation. Transformer-based bidirectional models [14, 20, 24, 31, 35] utilize full spatiotemporal attention to generate realistic and temporally coherent sequences. However, their bidirectional denoising requires access to the full sequences, limiting their use in interactive scenarios. To support causal prediction and long-horizon rollouts, autoregressive video models have been introduced [5, 19, 40, 43]. These methods generate frames sequentially in a manner more consistent with real-world dynamics. These advances in video generation have motivated the development of world simulators, whose goal is to predict the visual consequences of actions given the current state [39]. Recent advancements [7, 8, 10, 13, 21, 23, 26, 37, 38, 41, 42] illustrate how actions can be applied to guide visual outcomes. However, most of these existing approaches rely on coarse action vocabularies such as keyboard and mouse inputs or raw camera poses, which describe scene-level information adequately but do not enable dexterous hand–object interactions. This highlights the need for fine-grained embodied control signals in interactive egocentric video generation.

### 2.2. Camera- and Hand-conditioned Generation

In generated virtual environments, camera and hand motions jointly determine how people perceive and interact with their surroundings, making both modalities essential control signals for egocentric world simulators. Camera-conditioned video generation has been extensively explored with various condition-injection strategies [2, 3, 16, 17, 36]. For instance, ReCamMaster [3] injects camera extrinsic parameters through a dedicated camera encoder; CameraCtrl2 [17] encodes Plücker rays and adds them element-wise to visual features before the DiT module; and AC3D [2] adopts a more dynamic design by introducing camera embeddings via a ControlNet-style feedback branch. In contrast, hand-conditioned video generation remains relatively underexplored. PlayerOne [33] adds body pose embed-

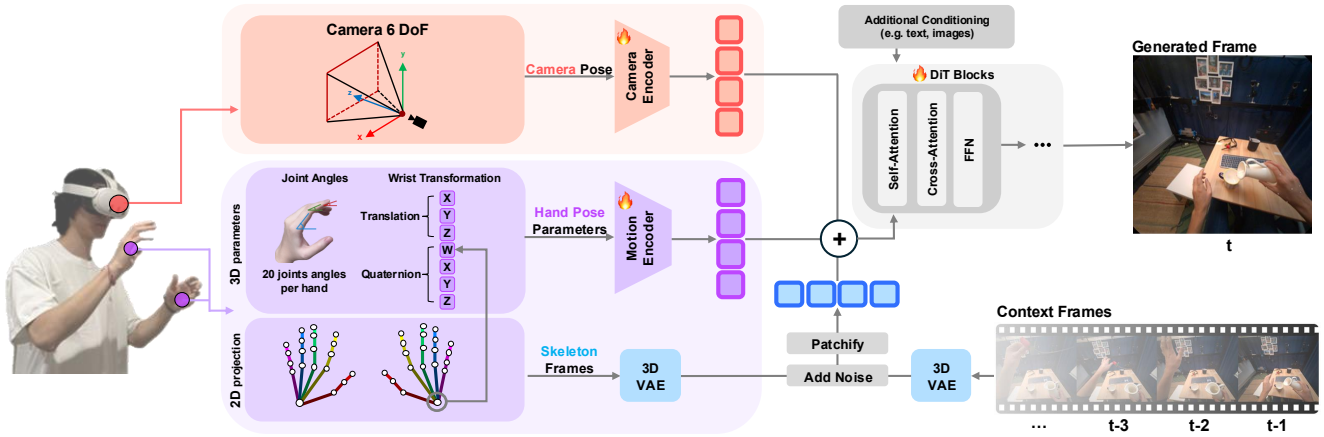



Figure 3. **Pipeline of generated reality system.** We track the head and hand poses of the user with a commercial headset. Hands are represented using the UmeTrack hand model [15], which includes translation and rotation of the wrist as well as rotation angles for 20 finger joints per hand. Our conditioning strategy employs a hybrid 2D–3D mechanism, combining a 2D image of the rendered hand skeleton (purple box, bottom) and the 3D model parameters (purple box, top). Features extracted from these modules are combined with the head pose features via token addition and fed into the diffusion transformer (DiT). The diffusion model autoregressively generates new frames at time $t$ using the last few generated frames as context in addition to the user-tracked conditioning signals.

dings to visual tokens before the DiT backbone, while PEVA [4] extends adaptive layer normalization (AdaLN) to inject pose information. However, both methods treat hands merely as part of the full-body pose, thereby limiting the granularity of hand control. InterDyn [1] employs binary masks instead of pose parameters as conditioning signals, which, however, increases the ambiguity between hand size and depth. In this work, we systematically compare various joint-level hand-conditioning strategies and identify a novel hybrid 2D–3D strategy that outperforms baselines on relevant metrics. We then incorporate this strategy into a camera-controlled video generation model, distill it into an autoregressive video generator, and evaluate this with users in an immersive format.

## 3. Conditional Video Generation with Tracked Head and Hands

In this section, we briefly review preliminaries on video diffusion models (Sec. 3.1). We then discuss hand pose representations and video model conditioning strategies, proposing a novel hybrid 2D–3D conditioning strategy (Sec. 3.2). We describe how to extend this framework to jointly condition on tracked head/camera poses, as well as joint-level hand signals (Sec. 3.3).

### 3.1. Preliminaries

Our study builds upon the Wan family of video generation models [35], a latent video diffusion transformer capable of generating temporally coherent video from a single input image or text prompt. The model consists of a 3D variational autoencoder $(\mathcal{E}, \mathcal{D})$ and a transformer-based diffusion model parameterized by $\Theta$. Given an input latent $z_0 = \mathcal{E}(V_0)$, the forward process follows the rectified flow formulation [9], where the noised latent is generated by linear interpolation:

$$z_t = (1-t)\, z_0 + t\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

with timestep $t \in [0, 1]$. The denoising process learns a velocity field $v_\Theta(z_t, t)$ that guides the transformation of noise back to data. The model is trained using a conditional flow matching [22], with objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, z_0, \epsilon}\left[\left\| v_\Theta(z_t, t) - u_t(z_0 \mid \epsilon)\right\|_2^2\right] \tag{2}$$

where $u_t$ is the target velocity derived analytically from the forward process. At inference, a sequence of latent frames is recovered by integrating $v_\Theta$ over time.

In the image-to-video (I2V) setting, the model is conditioned on an initial image $I_0$ encoded as $z_{\text{img}} = \mathcal{E}(I_0)$. The transformer-based denoiser $\mathcal{F}_\Theta$ autoregressively predicts video latents $\{z^{(f)}\}_{f=1}^{F}$, starting from $z_{\text{img}}$ and producing temporally consistent sequences. In the text-to-video (T2V) setting, the model is instead conditioned on a text prompt $p$ encoded as $z_{\text{text}} = \mathcal{T}(p)$ and starts the autoregressive generation from noise. The final video is reconstructed as $\hat{V} = \mathcal{D}(z^{(1)}, \ldots, z^{(F)})$.

### 3.2. Hand Pose-conditioned Video Generation

Conditioning strategies for video diffusion models have been widely explored, yet joint-level hand poses remain a challenging modality due to their high dimensionality and

complex articulation. We systematically study how to integrate hand poses into video diffusion transformers (DiT), focusing on two design choices: (1) the hand pose **representation**, i.e., how to represent tracked user hands, and (2) the **conditioning strategy**, i.e., how the conditioning information is injected into the generative model.

**Hand Pose Representation.** One option for the hand pose representation is a ControlNet-style pose video [44]. This representation is essentially a sequence of images that visualize the positions of human body joints and corresponding bones in the 2D pixel image space. In the context of egocentric hand conditioning, the video encodes a 2D hand skeleton rendered from the user's viewpoint.

While a skeleton video representation serves as a control signal spatially aligned with the image space, it inherently lacks 3D information. In immersive applications, a hand pose representation with 3D information is crucial for interactive video generation: in isolation, a 2D skeleton video exhibits depth ambiguity and suffers from self-occlusion as overlapping components of the skeleton make the position of certain hand joints ambiguous.

A 3D-aware hand representation is required for dexterous manipulation without ambiguities. Relevant parametric hand models are well known [15, 30] and usually model a hand pose as a 6 degree-of-freedom (DoF) transformation of the wrist along with rotation angles of each finger joint. We refer to the wrist pose and local joint rotations collectively as hand pose parameters (HPP). Applying standard forward kinematics to the HPP analytically yields the full set of 3D poses for all joints.

For compatibility with our training data [6], we adopt the UmeTrack hand model, whose HPP consist of 20 joint angles describing hand articulation together with the wrist pose. These HPP provide metric precision in depth and hand articulation, complementing the coarse but spatially grounded skeleton video representation.

**Hand Pose Conditioning.** To effectively incorporate hand pose parameters into the generative backbone, we examine four widely used condition injection strategies: (1) *token concatenation*, (2) *token addition*, (3) *adaptive layer normalization (AdaLN)*, and (4) *cross-attention fusion*. A pretrained variational autoencoder with encoder $\mathcal{E}$ projects a hand-contained raw video $V_r$ into the latent space, $z_r = \mathcal{E}(V_r)$, where $z_r \in \mathbb{R}^{b\times f\times c\times h\times w}$ is the latent of the raw video and $b$, $f$, $c$, and $h \times w$ denote batch size, frame count, channel dimension, and spatial size, respectively. We additionally extract the hand pose parameters of the same video, denoted as $H \in \mathbb{R}^{b\times f\times d}$, where $d$ is the dimensionality of the HPP. For token concatenation (1), we add additional input channels to the input convolutional layer and concatenate the embedded HPP features with the video latents along the channel dimension before patchification:

$$x = \text{patchify}([\ z_r,\ \mathcal{E}_{\text{conv}}(H)\ ]_{\text{channel-dim}}) \tag{3}$$

where $\mathcal{E}_{\text{conv}}$ denotes a lightweight motion encoder composed of 1D convolutional layers. For token addition (2), conditioning is applied through element-wise addition of HPP embeddings to patch tokens:

$$x = \text{patchify}(z_r)\ +\ \mathcal{E}_{\text{conv}}(H) \tag{4}$$

For AdaLN (3), the hand features modulate the activations within each DiT block through adaptive scale and shift vectors, a method inspired by adaptive normalization in conditional transformers [27]:

$$x = \alpha(H) \odot v_r + \beta(H) \tag{5}$$

where $\alpha(H)$ and $\beta(H)$ are learned from $H$, and $\odot$ denotes the Hadamard product. Finally, for cross-attention fusion (4), HPP embeddings serve as keys and values in motion-conditioned cross-attention layers injected after selected Transformer blocks, following the after-block cross-attention design of recent works [12]:

$$x^{(l+1)} = x^{(l)} + \text{CrossAttn}\big(x^{(l)},\ \mathcal{E}_{\text{conv}}(H)\big) \tag{6}$$

**Hybrid 2D–3D Hand Pose Conditioning.** We propose a hybrid conditioning scheme that combines ControlNet-style 2D skeleton videos with the 3D-aware HPP. This strategy combines the efficiency of ControlNet with the spatial awareness of HPP. As shown in Sec. 4, *token addition* yields the best performance among the evaluated pose injection approaches. We therefore incorporate HPP into the skeleton-based video control branch via element-wise token addition. Specifically, a hand-contained raw video $V_r$ and its corresponding skeleton video $V_c$ are encoded by the same VAE encoder $\mathcal{E}$ to obtain $z_r$ and $z_c$, respectively. We then concatenate the two latents in a channel-wise manner, and inject the HPP features using token addition:

$$x = \text{patchify}([\ z_r,\ z_c\ ]_{\text{channel-dim}})\ +\ \mathcal{E}_{\text{conv}}(H) \tag{7}$$

This design allows the model to resolve depth and self-occlusion ambiguity while maintaining strong spatial grounding from the skeleton representation.

### 3.3. Joint Camera and Hand Control

In head-mounted display (HMD) formats, visual content must be generated dynamically based on user interaction. Therefore, the user's viewpoint (camera), left hand, and right hand are foundational control signals for interactive video generation. Hand interaction enables intent-driven movement of generated objects, and viewpoint interaction enables the user to view the generated content from new

perspectives. To support these interactions, we introduce a framework for **joint hand and camera conditioning**, enabling realistic egocentric video generation driven by natural user interactions.

**Camera Pose Representation.** Previous works on pose-conditioned video generation often infer camera poses implicitly from body kinematics. For example, PlayerOne [33] estimates rotation-only camera trajectories from head pose with exocentric videos, while PEVA [4] models viewpoint change via body joint signals without explicitly modeling camera extrinsics. In contrast, we directly exploit the built-in inertial sensors and egocentric cameras of modern HMD, which provide a 6-DoF camera pose in world space, including both rotation ($r \in \mathbb{R}^{3\times3}$) and translation ($t \in \mathbb{R}^3$). This explicit camera representation enables the accurate modeling of the camera (or head) pose, making the generated video responsive to a user's head motion.

**Joint Conditioning Strategy.** We transform the 6-DoF camera poses into per-frame Plücker embeddings $P \in \mathbb{R}^{b\times f\times 6\times h\times w}$ [32], which are then projected into the same shape as the patch tokens with encoder $\mathcal{E}_{\text{cam}}$. We then apply element-wise addition over three components in the latent space: (a) video latents, (b) HPP embeddings, and (c) camera embeddings:

$$x = \text{patchify}([\, z_r,\, z_c \,]_{\text{channel-dim}}) + \mathcal{E}_{\text{conv}}(H) + \mathcal{E}_{\text{cam}}(P) \tag{8}$$

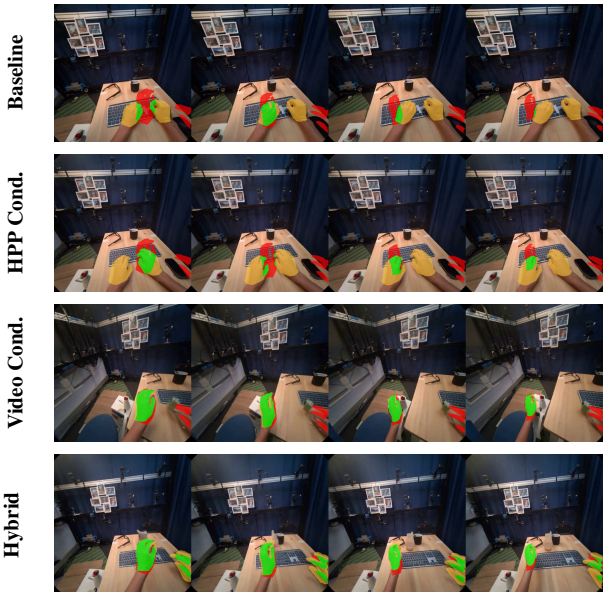
The fused representation $x$ is then passed into the DiT blocks for generation. During training, both hand and camera signals are jointly optimized under a unified conditioning schema, ensuring coherent motion alignment between user actions and egocentric viewpoint changes. An overview of this joint conditioning architecture is shown in Figure 3.

**Iterative Encoder Training.** In practice, we find jointly training both encoders from scratch to be unstable. We attribute this to (1) both camera and HPP embeddings being added in the same operation and (2) ambiguity between motion caused by hand interaction and camera movement. Thus, we adopt an iterative training approach: camera and HPP encoders are first trained independently, with the camera encoder weights initialized from the FUN model [35]. Then, both encoders are trained jointly in a final fine-tuning step to merge the conditionings.

## 4. Experiments

**Implementation details.** Building upon the Wan2.2 14B image-to-video (I2V) generation model [35], we first conduct a systematic study to determine the most effective hand motion conditioning strategy. Experiments are performed on the HOT3D dataset [6], which captures hand–object interactions with precise 3D hand annotations obtained via optical-marker motion capture and synchronized camera pose annotations. We segment each video into 5-second clips, yielding 5824 training samples, and reserve an unseen sequence of 45 clips for evaluation. For each of the conditioning strategies described in Sec. 3.2, we train LoRA [18] modules with rank 32 on both low-noise and high-noise experts for over 1K steps at a resolution of $480 \times 480$, using a learning rate of $1 \times 10^{-5}$ and a batch size of 16.

Figure 4. **Qualitative comparison of hand-pose conditioning strategies.** Ground-truth conditioning hand input is shown in red. Predicted hands are orange; overlap is green. Our hybrid conditioning strategy is most accurate among these baselines, especially when hands are partly occluded at the boundaries of the frame.




**Metrics.** We evaluate our model along three dimensions: overall video quality, hand pose accuracy, and camera pose accuracy. For video quality, we report PSNR for pixel-level accuracy, LPIPS [45] for perceptual similarity, SSIM for structural consistency, and Fréchet Video Distance (FVD) for distribution-level realism. For hand pose accuracy, we use WiLoR [28] to evaluate Procrustes Aligned Mean Per-Joint Position Error (PA-MPJPE) computed over 20 joints to measure 3D pose accuracy, and Procrustes Aligned Mean Per-Vertex Position Error (PA-MPVPE) computed over 778 vertices to measure 3D hand shape accuracy. We further compute the average L2 distance between ground truth and generated hand landmarks in the pixel space of each 2D frame [29]. Camera pose accuracy is evaluated by ex-

Table 1. **Quantitative comparison of hand-motion conditioning strategies.** We perform an ablation study on the Wan2.2 14B model, evaluating hand pose parameters (HPP), binary mask, skeleton video, and hybrid conditioning schemes. Results are reported for both video quality as well as 3D and 2D hand pose accuracy, where best results are highlighted as **first** and second. Our hybrid strategy using both 2D skeleton projection and 3D HPPs achieves the best accuracy while maintaining a competitive video quality. Note that the position errors here are in millimeters and Procrustes aligned. ControlNet* represents the use of pixel-level image conditioning, but we do not copy the DiT blocks as done in the original ControlNet implementation.

| Method | | Video Quality | | | | Hand Pose Accuracy | | |
|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | LPIPS↓ | SSIM↑ | FVD↓ | MPJPE↓ | MPVPE↓ | L2Err↓ |
| No Cond. | Baseline (Wan 2.2 Video 14B) | 14.59 | 0.4872 | 0.4855 | 601.55 | 17.86 | 12.29 | 67.50 |
| HPP Cond. | TokenConcat (PlayerOne [33]) | 15.09 | 0.4633 | 0.4983 | 560.34 | 18.02 | 12.34 | 65.43 |
| | AdaLN (PEVA [4]) | 15.02 | 0.4591 | 0.4906 | 677.26 | 18.49 | 12.53 | 65.97 |
| | CrossAttention | 14.71 | 0.4686 | 0.4840 | 662.22 | 17.56 | 12.04 | 63.23 |
| | TokenAddition (ReCamMaster [3]) | 15.19 | 0.4520 | 0.4975 | 601.15 | 17.84 | 12.14 | 56.66 |
| Video Cond. | Binary Mask (InterDyn [1]) | 16.58 | 0.3947 | 0.5533 | **356.11** | 12.83 | 9.56 | 35.64 |
| | Skeleton Video (ControlNet* [44]) | **16.89** | **0.3837** | **0.5601** | 389.26 | 12.38 | 9.25 | 11.72 |
| Hybrid Cond. | Skeleton Video + HPP Cond. | 16.85 | 0.3874 | 0.5574 | 383.69 | **12.23** | **9.10** | **11.50** |

Table 2. **Quantitative comparison of joint hand and camera conditioning strategies.** Compared with the camera-only and hand-only baselines, JointCtrl achieves the best overall performance across video quality, hand pose, and camera pose metrics. It maintains the highest visual quality while delivering competitive control accuracy for both hand and camera signals, relative to models specialized in a single modality. Translation and rotation errors are reported in meters and degrees, respectively.

| Method | | Video Quality | | | | Hand Pose Accuracy | | | Camera Pose Accuracy | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | LPIPS↓ | SSIM↑ | FVD↓ | MPJPE↓ | MPVPE↓ | L2Err↓ | TransErr↓ | RotErr↓ |
| CamCtrl | CameraCtrl [16] | 18.58 | 0.2943 | 0.6099 | 558.94 | 18.37 | 12.72 | 50.33 | **0.23** | **2.77** |
| HandCtrl | Best in Tab. 1 | 16.85 | 0.3874 | 0.5574 | **383.69** | **12.23** | **9.10** | **11.50** | 2.27 | 13.40 |
| JointCtrl | Ours | **18.60** | **0.2800** | **0.6173** | 396.93 | 12.81 | 9.66 | 13.42 | 0.25 | 2.79 |

tracting estimated trajectories from generated clips using GLOMAP [25] and computing rotation error (RotErr) and translation error (TransErr) following previous work [3].

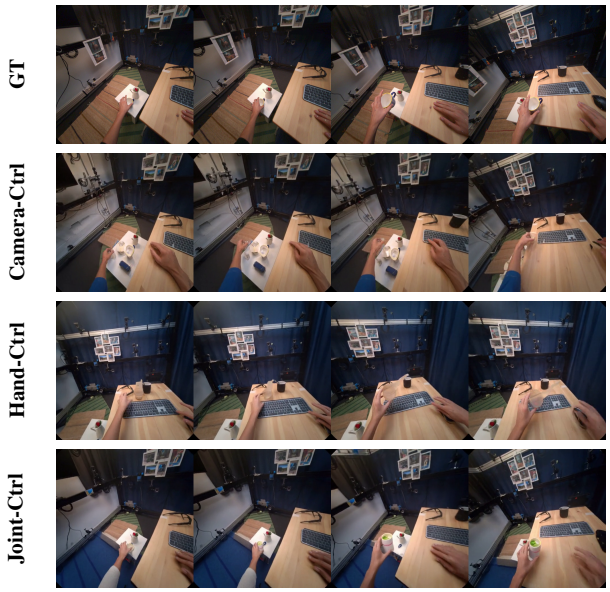
**Evaluating Hand-pose Conditioning.** Among the four injection strategies evaluated for conditioning on hand pose parameters (HPP), the *token addition* method achieves the best performance across hand pose accuracy metrics, as shown in Table 1. In contrast, *cross-attention* and *AdaLN* struggle to establish a stable mapping between HPP and visual features, likely due to the limited scale of the HOT3D dataset and the high dimensionality of the HPP, performing worse than the unconditioned baseline.

We further evaluate hybrid conditioning that integrates both skeleton video and HPP information. As shown in Table 1, the hybrid approach achieves the best performance across all hand accuracy metrics. Although the numerical gains over the ControlNet-style 2D skeleton-image conditioning strategy are moderate, likely due to the relatively simple hand motions in HOT3D, the hybrid 2D–3D method still produces more stable and anatomically faithful hand reconstructions qualitatively.

To contextualize the quantitative results for hand pose accuracy, we estimate lower bounds for the different metrics by evaluating the HOT3D test annotations under the same protocol, i.e., by fitting a 3D hand model using WiLoR [28] to the ground truth test images and evaluating our hand pose accuracy metrics. This yields MPJPE of 9.42, MPVPE of 7.74, and an L2 landmark error of 9.08, representing the inherent accuracy and uncertainty of the WiLoR-based hand pose estimator we use for all generated frames. Table 1 (right) shows that our hybrid conditioning method approaches this lower bound. To further validate robustness beyond HOT3D, we evaluate on the larger GigaHands dataset [11] and observe consistent improvements over 2D-only conditioning (Appendix B.1).

Qualitative comparisons in Figure 4 highlight these improvements, where predicted hands are shown in orange, ground truth in red, and their overlap in green. In the challenging case shown in this figure, ControlNet conditioning fails to reconstruct hands near the image boundary due to incomplete skeleton inputs, whereas the hybrid model gener-

Figure 5. **Qualitative comparison of joint hand–camera control.** Ground-truth (GT), camera-only, hand-only, and joint-control results. Camera-Ctrl and Hand-Ctrl are effective at controlling one of these modalities but not the other. Our Joint-Ctrl mechanism enables simultaneous control of camera and hands.




ates complete and spatially consistent hand structures even when hands are close to the frame edge.

**Evaluating Joint Head- and Hand-pose Conditioning.** We compare the proposed joint hand–camera conditioning framework against hand-only (HandCtrl) and camera-only (CameraCtrl [16]) baselines. As shown in Table 2, the joint-control model achieves the best video quality and balanced performance across hand and camera pose metrics. Specifically, CameraCtrl achieves the lowest rotation and translation errors in camera pose but fails to maintain accurate hand alignment, whereas HandCtrl produces precise hand poses but lacks camera control. Our joint-control model bridges this gap, achieving coherent coordination between hand motion and head dynamics. Figure 5 further illustrates that, without camera control, the hand-only model often interacts with incorrect objects. In this example, the hand-only model incorrectly predicts user intent by reaching toward an object on the table instead of the cup on the left.

## 5. The Generated Reality System

Using our detailed analysis of joint-level hand- and head–conditioned video generation, we next develop our generated reality system. This is a variant of the aforementioned video diffusion model, rolled out in a causal, i.e., autoregressive, manner and distilled to achieve interactive frame rates. The user's head and hand poses are dynamically tracked with a commercial VR system and used to condition the video generation model, whose output is streamed directly to the headset worn by the user.

**Autoregressive Distillation.** Following the self-forcing strategy, we distill a bidirectional Wan2.2 5B teacher model that is trained with our head- and hand-conditioning strategy into a causal 5B student model [19]. Autoregressive videos are generated in 12-frame chunks, complete with per-frame hand and head conditioning as outlined. The model supports both image-to-video (I2V) and text-to-video (T2V) settings. The resulting system provides a closed-loop generative experience—users can continuously move their hands and head, and the model renders the corresponding virtual response.

**Integration with VR System.** A real-time generative VR system is implemented with Unity on the Meta Quest 3. We use the captured head and hand poses from the Quest as our conditioning. This conditioning is streamed to a server hosting the distilled autoregressive model. For each video chunk, conditionings are read from a circular frame buffer with the most recent tracked data. Generated video chunks are then streamed back to the Quest 3 for interactive viewing in VR. We achieve 11 FPS in real-time with 1.4 seconds of latency on a single H100 GPU. The latency is bottlenecked by the time to generate and decode a 12-frame chunk. The added conditioning adds only an additional 0.002 s of latency.

**User Study Design.** To evaluate our generated reality system, we conducted two user studies. For this purpose, we recruited 11 subjects (age range = 22–30 years). The cohort consisted of 4 female and 7 male participants; 6 of them wore glasses. All participants reported normal or corrected-to-normal vision.

We designed the three different environments shown in Figure 6 for our studies. Observing these in the Quest headset, we ask users to perform the following tasks: "push the green button", "open the jar", and "turn the steering wheel", respectively. Users had a total of 8 seconds to complete a task. We tested two conditions for each task: one using our hand- and head-pose conditioned model and one baseline model that uses only head-pose conditioning. The relative difference between these conditions, therefore, demonstrates the effectiveness of hand control in our application. The baseline relies purely on the text-conditioned video model to complete the task without the user directly controlling the generated rendering of their hands. Users completed each of the three tasks four times (twice for each of the conditions), all in random order. Before starting each run, we asked users to roughly align their hands with the input image; we overlay their real-time hand pose to assist

Figure 6. **User study tasks and setup**. Our subjects completed three tasks using a commercial virtual reality headset: "push the green button", "open the jar", and "turn the steering wheel". Representative screenshots of all three tasks from the perspective seen by the user (top). A photo of our setup, in which the generated video's hands reflect the user's in real-time (bottom).

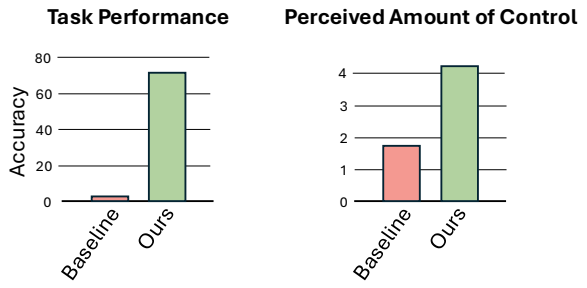



Figure 7. **User evaluation**. We show human subjects interactive videos without (baseline) and with (ours) tracked hand conditioning signals. For the baseline, we prompt the video model to complete the task using the same instructions provided to human subjects in our setting. In our tracked hand conditioning setting, users can more accurately complete the task than a video model can with just text conditioning (left). Moreover, users report a significantly higher level of perceived control over the interaction in our setting compared to the baseline, measured using a 7-point Likert scale (right).

with this process. Once they indicated alignment, we disabled the hand pose overlay and began the interactive experience, so users saw only the environment, the generated hands, and the results of hand-object interactions. Users were allowed two practice runs to familiarize themselves with the process before we began recording results. More details are outlined in Appendix C.

**Evaluating Task Efficiency.** As shown in Figure 7 (left), the baseline achieved an average of 3.0% for task accuracy, demonstrating that text prompts alone are insufficient for reliably completing tasks that require fine-grained hand–object interaction. Under identical conditions, our hand-controlled model achieved 71.2% task accuracy on average, highlighting the substantial improvement in task success provided by explicit hand controls.

**Evaluating User Experience.** After each trial, participants rated their perceived amount of control on a 7-point Likert scale (1 = worst, 7 = best). Shown in Figure 7 (right), our hand-controlled model received a mean score of 4.21, compared to 1.74 for the baseline. These results indicate that participants experienced markedly greater control over hand pose and movements with explicit hand conditioning than with text prompts alone, aligning with the observed improvements in task success.

## 6. Discussion

We present crucial first steps towards a vision of human-centric world simulation. Specifically, we identify and evaluate efficient and effective mechanisms for conditioning video diffusion models on tracked head and joint-level hand data. Moreover, we present a first version of an interactive generated reality system and demonstrate its efficacy with user studies.

**Limitations.** The resolution, latency, stereo rendering capabilities, image quality, and computing efficiency of our system lag far behind those of modern virtual reality systems. As with all current autoregressive video models, drift significantly degrades the image quality after a few seconds of rollout. Yet, the promise of generating an interactive and immersive virtual environment in a zero-shot manner is unprecedented and motivates future research on solving these issues.

**Future Work.** Improving the aforementioned limitations towards retinal image resolution in stereo with imperceptible (i.e., $< 20$ ms) latency and long rollouts on a wearable computer embedded in a headset is an enormous challenge. Yet, most of these problems are well aligned with ongoing research and development efforts on autoregressive video diffusion models across the computer vision and AI communities.

**Conclusion.** Generated reality could enable immersive learning and exploration, letting users acquire skills and practice complex tasks in a zero-shot manner, without the need for laborious modeling of 3D virtual environments.

## References


[1] Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12467–12479, 2025. 4, 7

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers, 2025. 3

[3] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video, 2025. 3, 7

[4] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction, 2025. 3, 4, 6, 7

[5] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025. 3

[6] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Introducing hot3d: An egocentric dataset for 3d hand and object tracking, 2024. 5, 6

[7] Junhao Cheng, Yuying Ge, Yixiao Ge, Jing Liao, and Ying Shan. Animegamer: Infinite anime life simulation with next game state prediction, 2025. 3

[8] Etched Decart, Q McIntyre, S Campbell, Xinlei Chen, and R Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2024. 3

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 4

[10] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control, 2024. 3

[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, 2025. 7, 13

[12] Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025. 5

[13] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft, 2025. 3

[14] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. 3

[15] Shangchen Han, Po-Chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 Conference Papers*. Association for Computing Machinery, 2022. 4, 5

[16] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 7, 8

[17] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models, 2025. 3

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 6

[19] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. 3, 8

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou,

Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. 3

[21] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N. Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image, 2025. 3

[22] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 4

[23] NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. 3

[24] OpenAI. Sora: Creating video from text, 2024. 3

[25] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited, 2024. 7, 13

[26] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 3

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 5

[28] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2025. 6, 7, 13

[29] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images, 2024. 6

[30] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 2017. 5

[31] Abhishek Sharma, Adams Wei Yu, Ali Razavi, Andeep Toor, Andrew Pierson, Ankush Gupta, Austin Waters, Aäron van den Oord, Daniel Tanis, Dumitru Erhan, Eric Lau, Eleni Shaw, Gabe Barth-Maron, Greg Shaw, Han Zhang, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jakob Bauer, Jeff Donahue, Junyoung Chung, Kory Mathewson, Kurtis David, Lasse Espeholt, Marc van Zee, Matt McGill, Medhini Narasimhan, Miaosen Wang, Mikołaj Bińkowski, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Nando de Freitas, Nick Pezzotti, Pieter-Jan Kindermans, Poorva Rane, Rachel Hornung, Robert Riachi, Ruben Villegas, Rui Qian, Sander Dieleman, Serena Zhang, Serkan Cabi, Shixin Luo, Shlomi Fruchter, Signe Nørly, Srivatsan Srinivasan, Tobias Pfaff, Tom Hume, Vikas Verma, Weizhe Hua, William Zhu, Xinchen Yan, Xinyu Wang, Yelin Kim, Yuqing Du, and Yutian Chen. Veo: a text-to-video generation system. Technical report, Google DeepMind, 2025. 3

[32] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering, 2022. 6

[33] Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator, 2025. 3, 6, 7

[34] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines, 2024. 3

[35] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 4, 6

[36] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[37] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[38] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory, 2025. 3

[39] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators, 2024. 3

[40] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models, 2025. 3

[41] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[42] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025. 3

[43] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation, 2025. 3

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 5, 7

[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 6

# Generated Reality: Human-centric World Simulation using Interactive Video Generation with Hand and Camera Control

Supplementary Material

## A. Experiment Details

### A.1. Initialization of the Motion Encoder

Our experiments are based on the Wan2.2 14B model family, which uses a mixture-of-experts (MoE) architecture with two DiT experts: one specialized for high-noise steps and one for low-noise steps. To train the motion encoder effectively under this design, we adopt a continual training scheme.

During high-noise DiT training, we zero-initialize the motion encoder. After convergence, the trained encoder is transferred and used as the initialization for low-noise training. This two-stage setup provides a stronger starting point for the low-noise model and mitigates the impact of the limited HOT3D dataset, resulting in more stable training and improved motion alignment.

### A.2. Continual Training of DiT Experts

For hybrid conditioning, we aim to emphasize fine-grained alignment during training. To achieve this, we initialize the DiT with the LoRA weights learned from skeleton-video conditioning and continue training from this point. This provides the model with a well-structured spatial prior and allows the hybrid training stage to focus on refining articulation and depth cues introduced by the hand pose parameters.

Similarly, for joint hand–camera conditioning, we initialize the DiT with the LoRA weights obtained from the hybrid model and then train with both hand and camera inputs. This continual training strategy gives the joint model a strong initialization and leads to more stable convergence and improved motion consistency. Furthermore, it helps the model decouple the conditionings, which are both applied in the same token addition operation.

### A.3. Lower bounds

We estimate the lower bound of our evaluation pipeline by running the same metrics on the HOT3D validation annotations themselves. Hand poses are obtained from WiLoR [28], and camera trajectories are computed using GLOMAP [25]. This provides the inherent error level of the annotation and reconstruction process under our evaluation protocol.

## B. Additional Evaluation

### B.1. Alternative Datasets

In addition to HOT3D, we evaluate our method on the larger GigaHands [11] dataset (8× larger than HOT3D) with the Wan2.2 5B model. As shown in Table 4, we continue to yield consistent improvements over the baselines; particularly, our 2D–3D hybrid conditioning outperforms 2D only conditioning, reducing MPJPE by 10%, MPVPE by 11%, and 2D error by 34%. These results indicate scalability to larger, more complex data and richer hand motions. Fig. 9 and 10 provide additional qualitative comparisons across four scenes from the GigaHands dataset.

Table 3. **Lower bound for hand and camera pose evaluation metrics.**

| MPJPE↓ | MPVPE↓ | L2Err↓ | TransErr↓ | RotErr↓ |
|---|---|---|---|---|
| 9.42 | 7.74 | 9.08 | 0.0191 | 0.44° |

Table 4. **GigaHands ablation.** Additional hand pose accuracy ablations with Wan2.2 5B, trained on the GigaHands dataset. Hybrid conditioning continues to improve over 2D-only conditioning as dataset scale increases.

| Method | MPJPE↓ | MPVPE↓ | L2Err↓ |
|---|---|---|---|
| Ground-truth | 16.41 | 11.03 | 59.38 |
| Baseline | 20.86 | 15.08 | 268.49 |
| 3D Cond. | 20.63 | 14.90 | 250.79 |
| 2D Cond. | <u>19.67</u> | <u>14.03</u> | <u>134.77</u> |
| Hybrid Cond. | **17.78** | **12.48** | **89.59** |

### B.2. Text-to-Video Generation

Despite being trained on videos from a controlled studio environment, our model is able to transfer its hand interaction capabilities to diverse scenes unseen in training. To demonstrate "human-centric" generation beyond HOT3D's controlled hand-object interactions, we conduct text-to-video generation across complex, dynamic scenarios (Fig. 2).

## C. User Study Details

Fig. 8 visualizes comparisons between baseline and our method, captured during the user study. We chose short, simple tasks to enable objective (binary) completion measures and to isolate controllability from generation complexity and long-horizon drift; this also reduces participant discomfort given the current latency.

After each recorded run, participants are asked the question: "On a scale from 1-7, with 1 being no control and 7 being full control, rate the perceived controllability of the

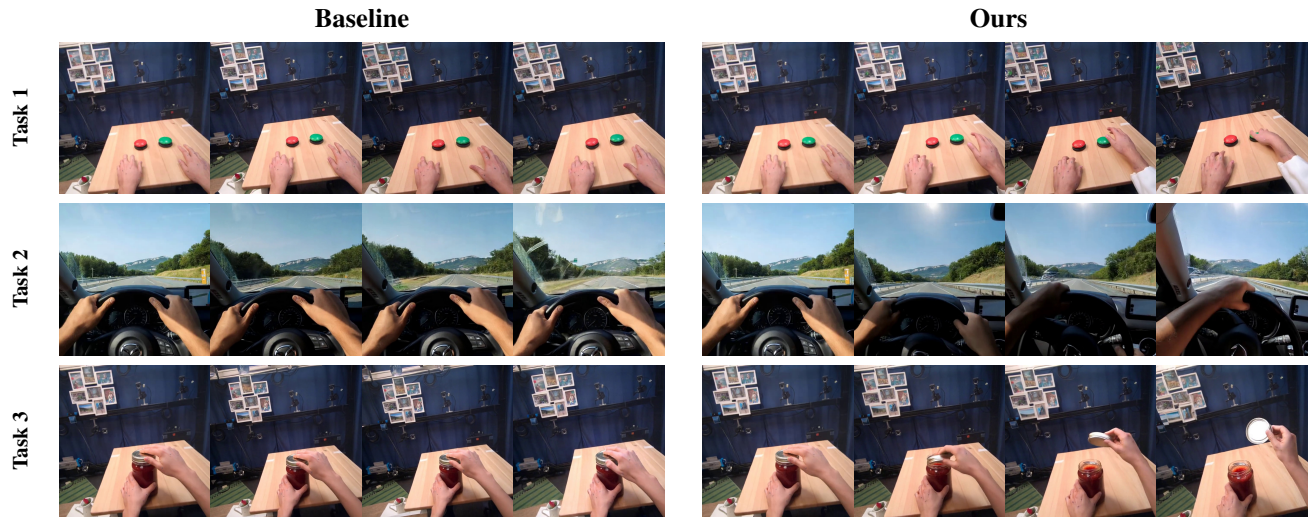



Figure 8. **User Study Qualitative Comparison.** Captured user study results of the baseline vs. our method.

system." To measure task completion, all generated videos from the session are blind-reviewed by a separate participant for a binary failure/success metric.

## D. Limitations

While the system models complex hand-object interactions, it struggles with longer-range hand-object-object dependencies. The causal model suffers drawbacks typical of DMD distillation methods, i.e., mode-seeking behavior and oversaturation over long horizons.

We acknowledge that 1.4 second latency is not sufficient for fully immersive XR systems. However, this latency is not fundamental to our approach and can be improved with better hardware, alternative distillation methods, and system optimization (e.g., we communicate with a remote GPU server rather than a local one). Despite this concern, we believe the system to be a practical tool for rapid prototyping and open-ended creation.

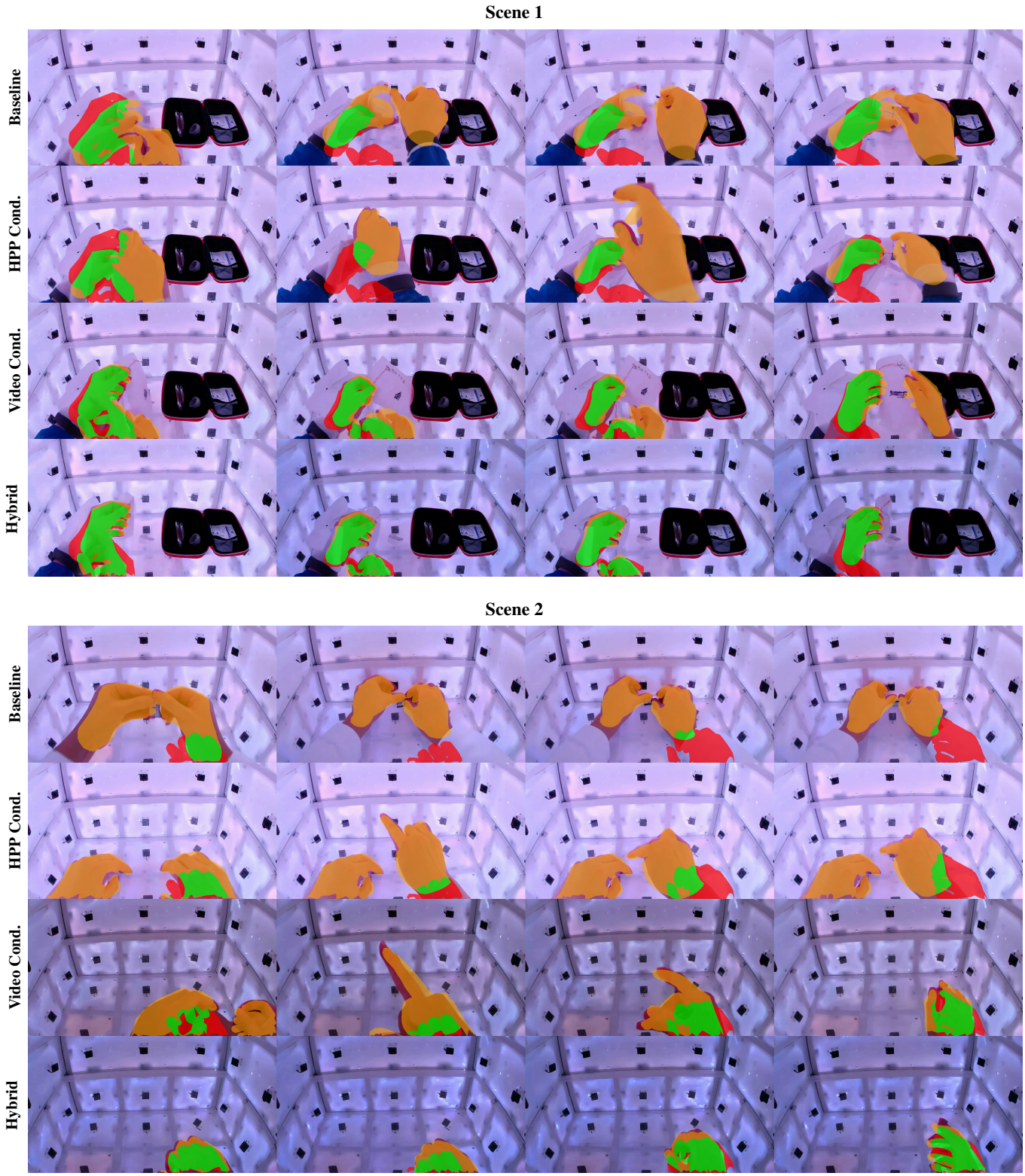



Figure 9. **GigaHands qualitative comparison (1/2).** Qualitative comparison of hand-pose conditioning strategies on the GigaHands dataset. Ground-truth conditioning hand input is shown in red. Predicted hands are orange; overlap is green. Our hybrid conditioning strategy continues to outperform.

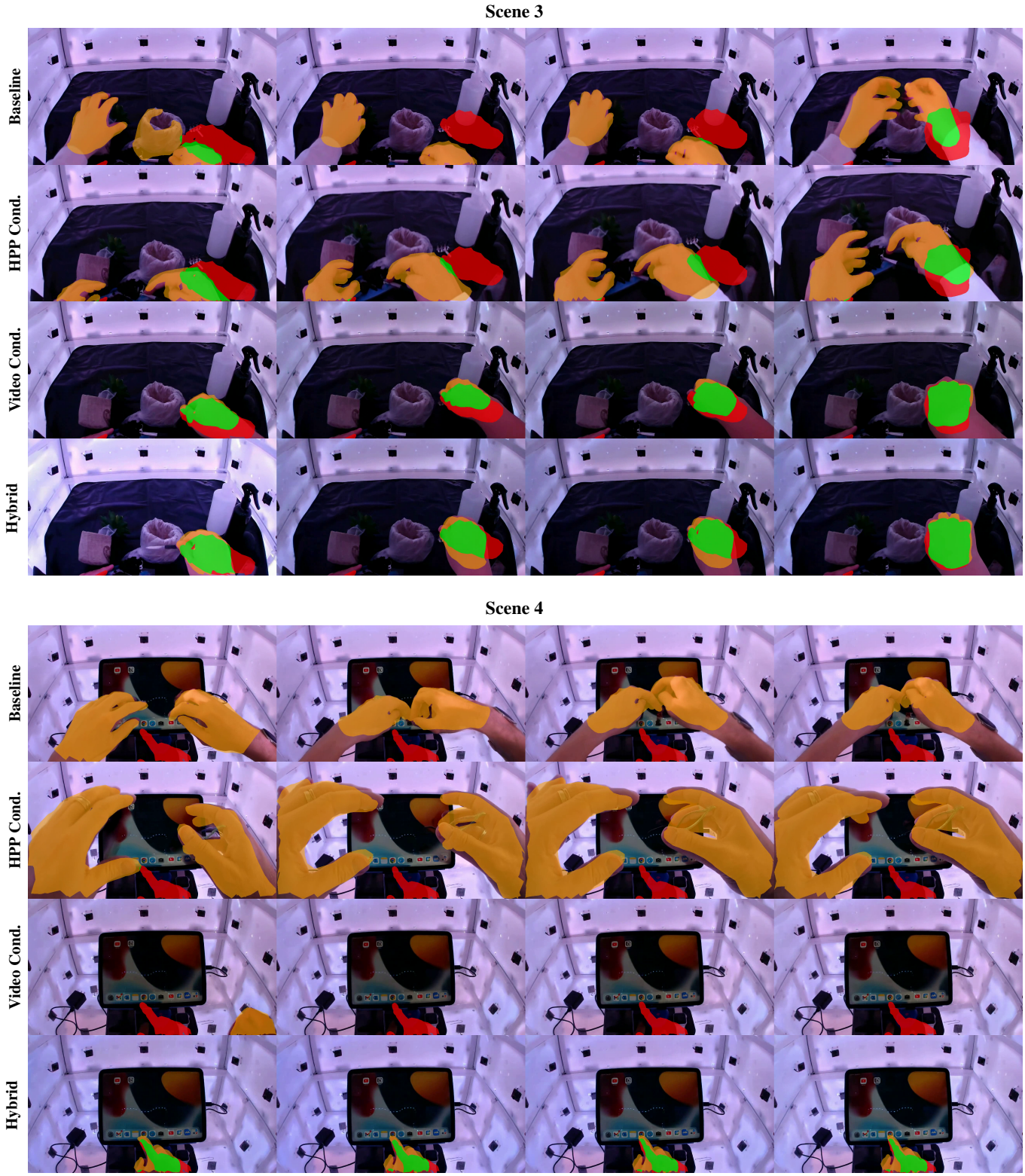



Figure 10. **GigaHands qualitative comparison (2/2).** Qualitative comparison continued. Ground-truth conditioning hand input is shown in red. Predicted hands are orange; overlap is green.