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 [3], and camera trajectories are computed using GLOMAP [2]. 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 [1] dataset (8× larger than HOT3D) with the Wan2.2 5B model. As shown in Table S2, 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. S3 and S4 provide additional qualitative comparisons across four scenes from the GigaHands dataset.

Table S1. **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 S2. **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. S2).

C. User Study Details

Fig. S1 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

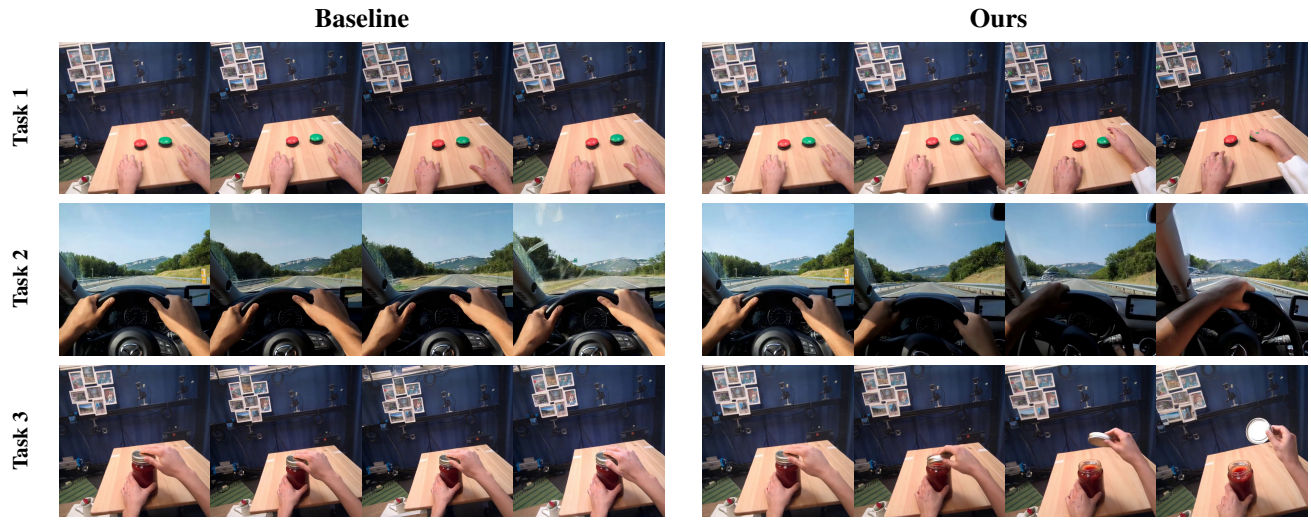


Figure S1. **User Study Qualitative Comparison.** Captured user study results of the baseline vs. our method.

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.

References

[1] 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. 1

[2] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited, 2024. 1

[3] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2025. 1

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...

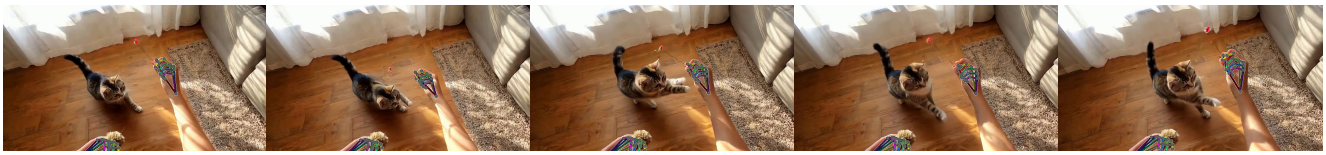
A gritty **medieval dungeon**... the right hand wields a **steel longsword**... an **armored soldier** charges towards the viewer...

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 S2. **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.

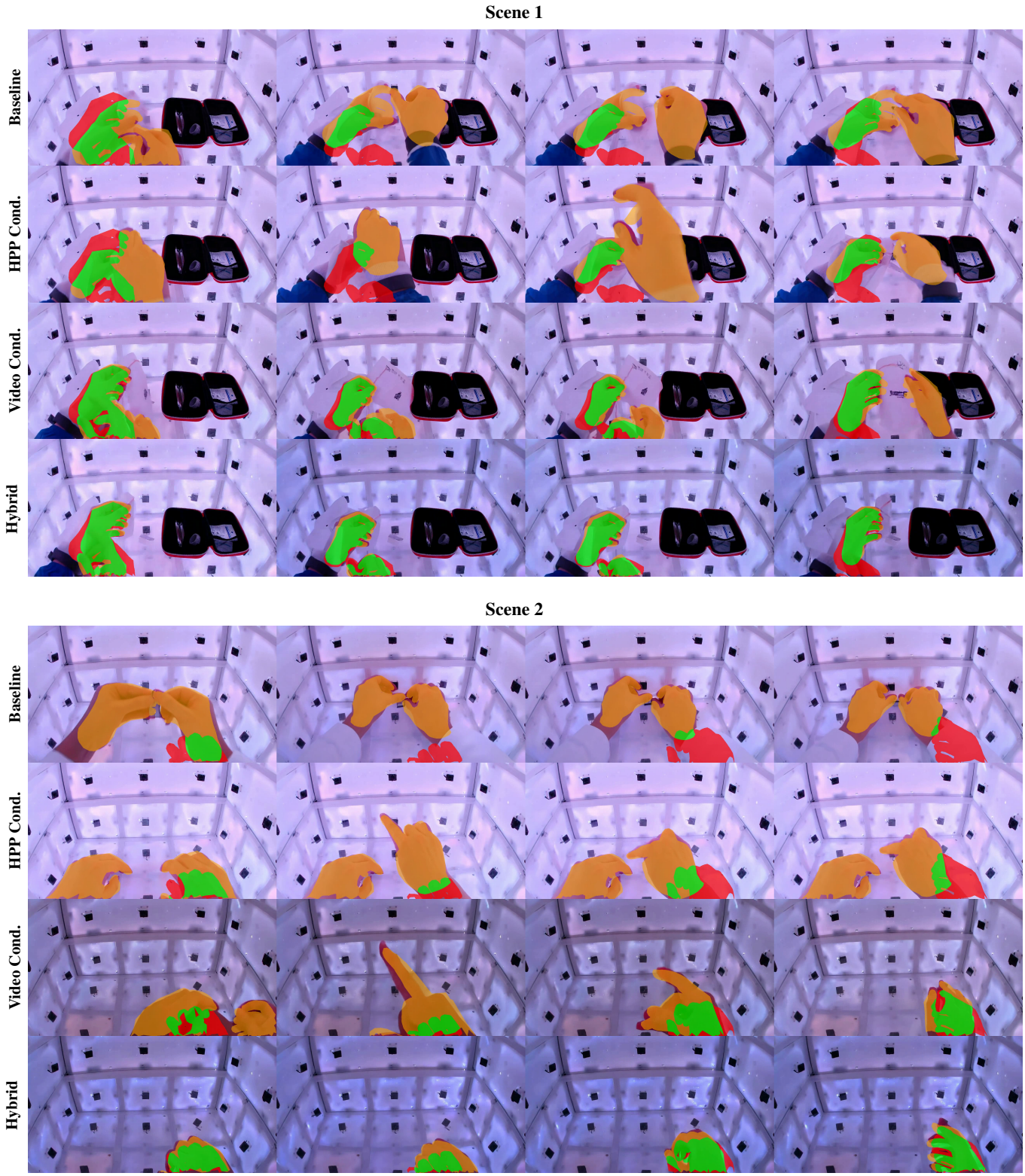


Figure S3. **GigaHands qualitative comparison (1/2).** Qualitative comparison of hand-pose conditioning strategies on the GigaHands dataset. Ground-truth conditioning hand input is shown in red. Predicted hands are orange; overlap is green. Our hybrid conditioning strategy continues to outperform.

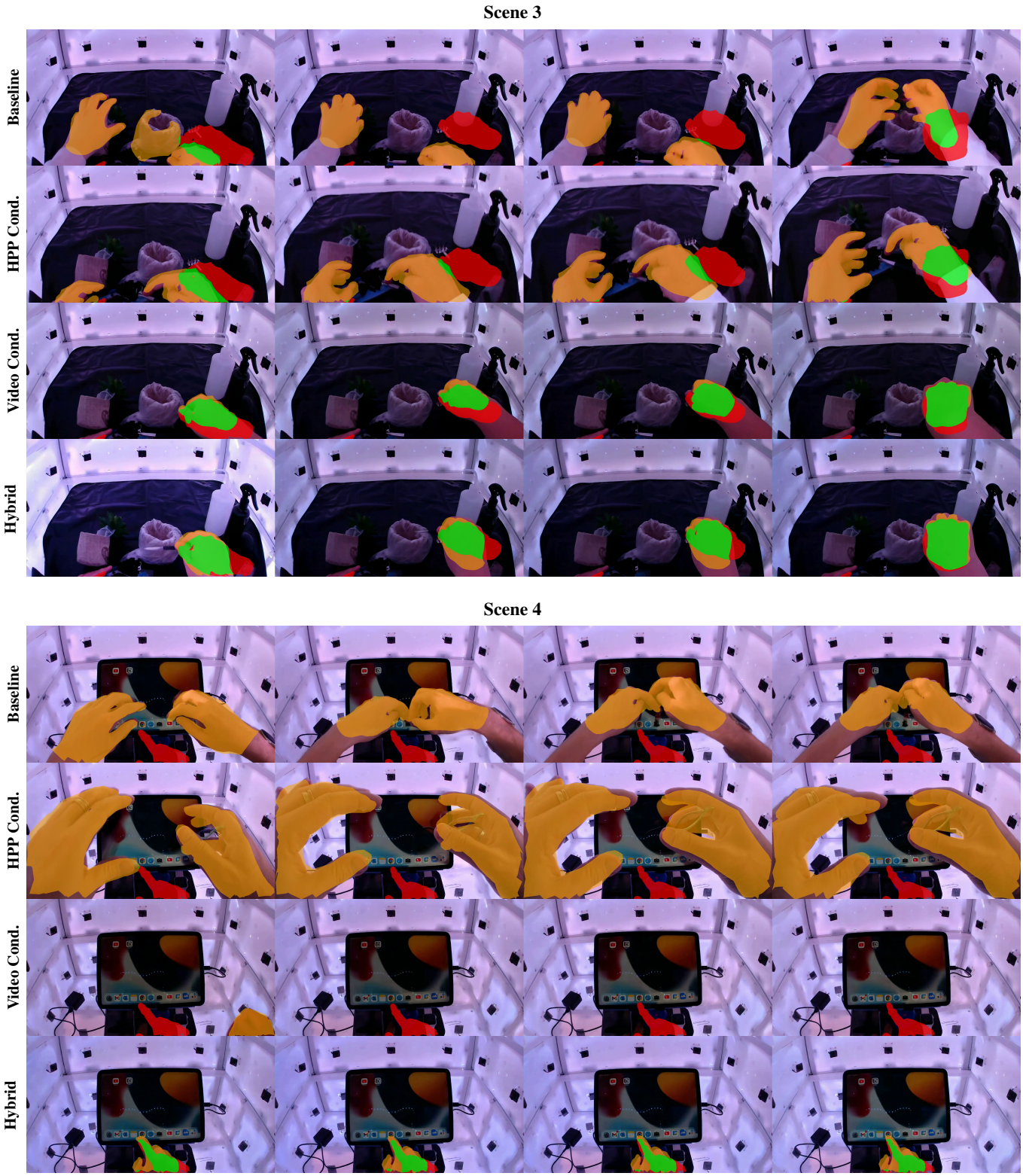


Figure S4. **GigaHands qualitative comparison (2/2).** Qualitative comparison continued. Ground-truth conditioning hand input is shown in red. Predicted hands are orange; overlap is green.