ASTER: UNIVERSAL WORLD SIMULATION VIA SCALABLE INTERACTION PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Interaction data offer a way to train video world simulators to predict the consequences of actions, but action representations differ across robots, human videos, and simulators. We present **ASTER**, a video world simulator conditioned on an initial frame, point flow on a moving body, and camera geometry. Only the mover's motion is specified; the surrounding scene response must be predicted. This shared input lets us pretrain an 8B video diffusion transformer on 13 real and simulated interaction sources. Across eight evaluation settings, **ASTER** reduces perceptual prediction error by 34% on average relative to the strongest external predictor and by 16% relative to its backbone, Cosmos 3. After adaptation to painting and laser cutting, it reaches the requested target in 83% of cases and produces a valid visible process in 73%, compared with 42% and 17% for a text-conditioned image-to-video baseline. At fixed model size, prediction improves across a 230× pretraining compute range, and the full robot, human, and object mixture lowers perceptual error relative to robot-only pretraining in all six evaluated source domains.

1 INTRODUCTION

Consider a paint roller moving against a blank wall. Given its motion, a video world simulator should predict where paint appears; changing that motion should produce a different letter (Figure 1). The same problem arises when a gripper lifts an object, a hand unties a knot, or a ball strikes a set of pins. In each case, the motion of one body is supplied, but the response of the objects and materials it affects must be predicted. Following the mover's trajectory is only part of the task.

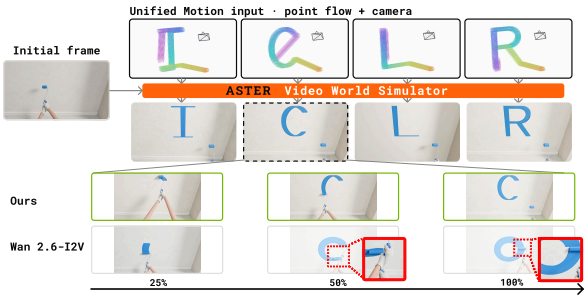

Figure 1: **One scene, different interactions.** Supplied roller motions produce I, C, L, and R. The lower rows compare how **ASTER** and Wan 2.6-I2V paint C over time.

Large video models generate realistic scenes and interactive environments (Yang et al., 2024; Ball et al., 2025; NVIDIA, 2026), yet a plausible image can conceal an incorrect interaction: a mark may appear before a tool reaches it, or an object may move without contact. Action-conditioned predictors offer more direct control (Zhu et al., 2025b; Guo et al., 2026), but their inputs often depend on a particular robot or action space. This makes it difficult to combine the available experience. Robot logs record joint commands or end-effector poses, human videos show hands and tools, and simulators provide object trajectories. These sources describe related interactions in different forms.

ASTER expresses these interactions through a shared geometric input: an initial image, point flow on the moving body, and camera geometry. We call the body whose motion is supplied the *mover*; it may be a gripper, hand, tool, or object. Its trajectory specifies the interaction without supplying the future state of the rest of the scene. We build on prior motion-based conditioning (Alzayer et al.,

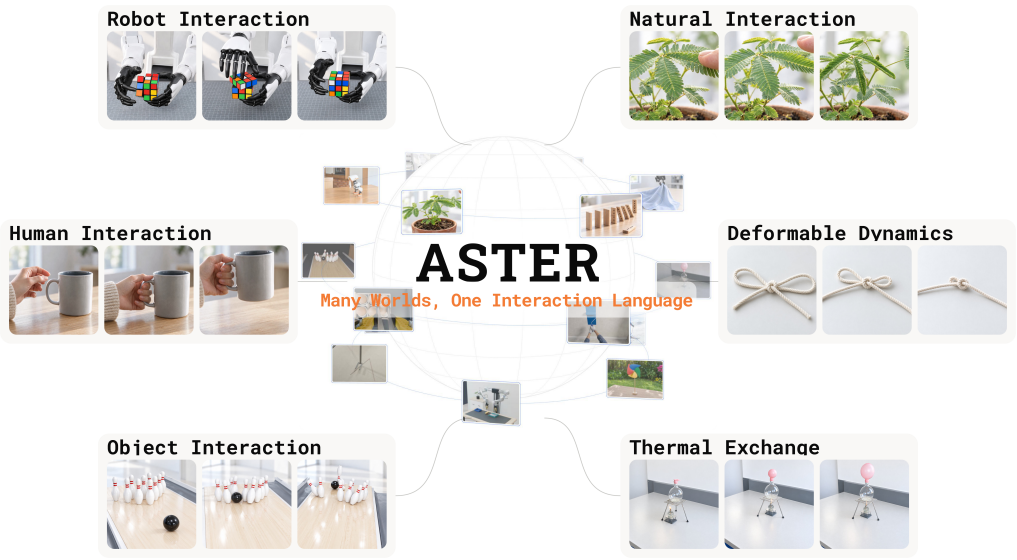

Figure 2: **One interaction language, diverse scene responses. ASTER** uses a common motion representation across robot, human, and object interactions to predict how the surrounding scene changes.

2026; Li et al., 2026c; Huang et al., 2026) to pretrain an 8B video diffusion transformer initialized from Cosmos 3 (NVIDIA, 2026) on 13 real and simulated sources. Figure 2 illustrates the range of interactions, and Figure 3 shows how motion and camera features condition the predictor.

Our experiments address four questions:

- **What does the model predict beyond the supplied motion?** Post-training examples show object transport, leaf folding, knot untying, and collision responses, with the affected scene elements left for the model to predict (Figure 4).
- **Does interaction pretraining improve prediction and transfer? ASTER** has the lowest LPIPS in all eight evaluation settings, including DROID after target adaptation. Average reductions are 34% relative to the strongest external predictor and 16% relative to Cosmos 3 (Table 1).
- **Can specified motions produce new outcomes?** After adaptation to painting and cutting, we test unseen targets and appearance changes. **ASTER** reaches the target in 83% of cases versus 42% for the text-conditioned baseline, and passes process validity in 73% versus 17%. Scoring both criteria distinguishes a correct final image from a consistent visible execution (Section 4.3).
- **How do compute and data composition affect performance?** At fixed model size, prediction improves over a 230× compute range. The full robot, human, and object mixture also lowers LPIPS relative to robot-only pretraining in all six source domains, including robot domains (Section 4.4).

2 Related Work

Video world models. World models predict the consequences of actions, from latent dynamics for planning and control (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025) to pixel-space simulators and interactive environments (Yang et al., 2024; Bruce et al., 2024; Ball et al., 2025). Video foundation models provide a starting point for learning such predictors (NVIDIA et al., 2025a; NVIDIA, 2026). Action-conditioned models use robot experience but are often designed around particular control interfaces (Zhu et al., 2025b; Guo et al., 2026; Zhu et al., 2025a). Counterfactual and structure-integrating models instead probe physical dynamics through alternative predictions and

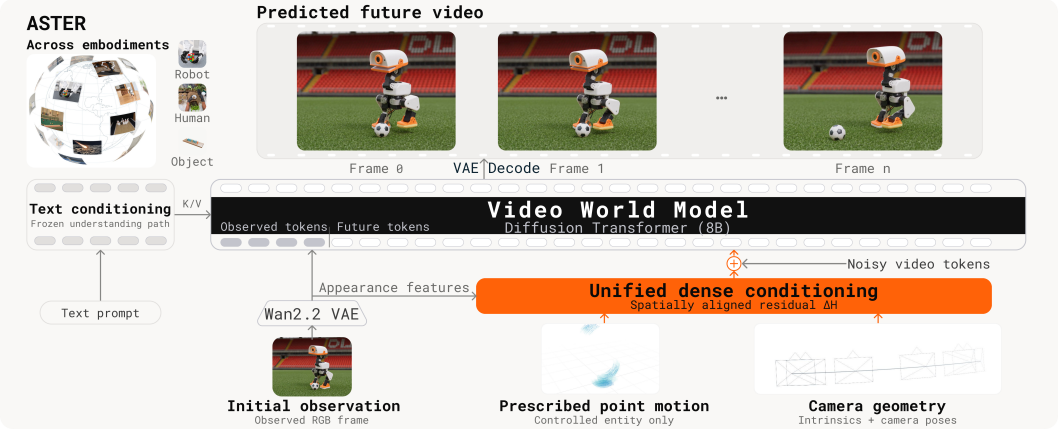

Figure 3: **ASTER overview.** Given an initial frame, the 3D point flow of the mover (a gripper, hand, tool, or object), and the camera trajectory, **ASTER** generates the future video. Appearance features sampled from the first frame are carried along the point flow and, together with camera-ray features, added to the video tokens of an 8B video diffusion transformer; text enters through a frozen understanding path. Robot, human, and simulated data all use this input format.

structured representations (Venkatesh et al., 2024; Kotar et al., 2025). Our focus is pooling interaction experience across sources with different action formats to improve a shared video predictor.

Motion and camera conditioning. Rendered embodiments, masks, and trajectories connect actions to visual prediction (Chen et al., 2026; Li et al., 2026a; Alzayer et al., 2026). Image-plane action flows and 3D point tracks support cross-embodiment prediction (Li et al., 2026c; Huang et al., 2026), while ray-based conditioning provides camera control (He et al., 2025; Gu et al., 2026). Trajectory-guided generation and editing also use prescribed motion to control video (Geng et al., 2025; Wang et al., 2025; Chu et al., 2025; Lee et al., 2026). **ASTER** builds on these motion-and-camera representations to pool heterogeneous interaction data. We study how joint pretraining improves prediction of the scene beyond the supplied mover trajectories.

Scaling and diverse interaction data. Compute and data allocation shape performance in language models (Kaplan et al., 2020; Hoffmann et al., 2022) and diffusion models (Liang et al., 2026; Yin et al., 2025; Chickering et al., 2026; Ge et al., 2026). Human videos, simulation, and data editing broaden the experience available to embodied models (Lepert et al., 2025a;b; 2026; Jang et al., 2025; Tian et al., 2025; Gao et al., 2026). We study interaction-pretraining compute and source composition at a fixed model size, including transfer beyond the pretraining domains. Because visual quality alone does not establish physical consistency (Motamed et al., 2026), we complement reference-video metrics with judgments of the visible process in specified painting and cutting tasks.

3 Method

ASTER predicts how a scene responds to a supplied motion. A shared geometric input lets the same model learn from robot commands, human motion, and simulated trajectories. Figure 3 shows the conditioning pathway.

3.1 Interaction representation

The inputs are an initial image I_0, which may contain multiple camera views, optional text c, and an interaction description $\mathcal{A} = (P_0, V, \mathcal{C})$. Here P_0 contains points on the moving body (the *mover*), V gives their interval velocities, and $\mathcal{C}$ specifies camera poses and intrinsics. Validity and point-view association masks accompany these quantities. The model predicts

$$p_\theta(I_{1:T} \mid I_0, \mathcal{A}, c). \tag{1}$$

Only the mover's motion is supplied. The positions, deformation, and appearance of affected objects remain prediction targets; no future scene image or target mask is provided.

Robot kinematics, hand and tool tracks, and simulated object states are converted to this format. Positions and camera translations use a shared per-source scale. Image-plane tracks remain 2D when calibrated geometry is unavailable, rather than being treated as metric 3D trajectories. Appendix A.1 gives the coordinate conventions and temporal alignment.

3.2 Motion and camera conditioning

We initialize an 8B video diffusion transformer from Cosmos 3 (NVIDIA, 2026) and retain its frozen video VAE. Initial-frame latents remain fixed during denoising, and text enters through the frozen understanding branch.

For each point and associated view, we sample an appearance descriptor from the initial-frame token grid. An MLP combines it with the initial position, cumulative and local displacements, normalized depth, and validity. These features are bilinearly splatted at the point's projected locations in future latent frames, forming $F_\tau(x)$ (Wang et al., 2025; Chu et al., 2025). Camera poses and intrinsics define Plücker-ray maps, which a spatiotemporal encoder maps to features $G_\tau(x)$ on the same grid (He et al., 2025); uncalibrated views receive zero rays. The video tokens receive the residual

$$H'_\tau(x) = H_\tau(x) + m_{\tau,x}\,\psi(F_\tau(x) + G_\tau(x))\,, \tag{2}$$

where m selects valid future tokens. Zero-initializing ψ preserves the pretrained predictor at initialization (Zhang et al., 2023). Appendix A specifies sampling, projection, and masking.

3.3 Training

We pretrain on 13 robot, human, and simulated sources (Appendix B) using a shared rectified-flow objective (Liu et al., 2023; Lipman et al., 2023). The model predicts video-latent velocity, excluding initial-frame and invalid-content positions from the loss (Equation 8). There is no auxiliary trajectory or segmentation loss. In multi-view training, we sometimes drop point-flow conditioning in selected views while predicting all views. For YAM-XDof, only the wrist cameras supply motion and camera conditioning; wrist and third-camera views are predicted jointly.

Training first updates the conditioning modules and generation-side key/value projections, then expands to the broader generation branch. The main run uses 128 GB200 nodes for 40,000 steps and 77,056.0 GPU-hours. Task adaptation reuses the same input format; Appendix E.2 describes its simulation and general-video data construction.

4 Experiments

We examine scene responses, prediction accuracy, specified-task execution, and the effects of compute and data composition.

Data and metrics. The benchmark uses held-out windows from six pretraining domains: Deform360 (Li et al., 2026b), MECKA, Behavior-1K (Li et al., 2024), YAM-XDof, GenieSim3.0 (Yin et al., 2026), and SGD-PhysX. We also evaluate adaptation to DROID (Khazatsky et al., 2024), which is excluded from pretraining. VGG-based LPIPS (Zhang et al., 2018), PSNR, and SSIM (Wang et al., 2004) compare generated future frames with recorded video, excluding the initial frame and invalid image content. The post-training showcases and specified-task study are separate from this benchmark.

Comparisons. Cosmos 3 shares the backbone but has no interaction pretraining; Robot-only uses robot-only pretraining. External predictors are Wan 2.6-I2V (Alibaba Cloud, 2025), Wan-Move (Chu et al., 2025), Cosmos-Predict2.5 (NVIDIA et al., 2025b), and Ctrl-World (Guo et al., 2026). They use their supported inputs and are not compute-matched to **ASTER**. Appendix C gives the benchmark protocols and uncertainty; Appendix E.1 describes the separately matched fine-tuning protocol for visual comparisons.

Table 1: **ASTER** gives the best prediction in all eight settings. LPIPS measures perceptual error (lower is better). Cosmos-Predict2.5 is the strongest external predictor by LPIPS; Cosmos 3 is our backbone without interaction pretraining. Reductions are relative to each reference, averaged across settings. Appendix Table 2 includes all predictors, PSNR, SSIM, and uncertainty.

	Perceptual error (LPIPS ↓)			Relative reduction	
Setting	Cosmos-Predict2.5	Cosmos 3	**ASTER**	vs. Cosmos-Predict2.5	vs. Cosmos 3
Deform360	0.3745	0.2774	**0.2204**	−41.1%	−20.5%
MECKA	0.5032	0.4026	**0.3910**	−22.3%	−2.9%
Behavior-1K	0.5395	0.3996	**0.3828**	−29.0%	−4.2%
YAM-XDof	0.3608	0.2886	**0.2012**	−44.2%	−30.3%
GenieSim3.0	0.5301	0.4078	**0.2133**	−59.8%	−47.7%
SGD-PhysX	0.3685	0.2948	**0.2766**	−24.9%	−6.2%
DROID (10%)	0.4111	0.3162	**0.3016**	−26.6%	−4.6%
DROID (100%)	0.3636	0.2909	**0.2640**	−27.4%	−9.2%
Average	0.4314	0.3347	**0.2814**	−34.4%	−15.7%

4.1 Predicting scene responses (RQ1)

Figure 4 shows predictions after task-focused post-training on simulation and general videos. Only the mover's trajectory is supplied: the model must predict the affected cap, egg, leaves, rope, pins, and particles. The examples span object transport, deformation, and collision responses, rather than simply reproducing the moving body's path.

Each row shows ordered frames from one source window. Motion is displayed separately or projected onto the initial image; these are two views of the same conditioning interface. The examples illustrate post-training behavior, not a task-success estimate. Appendix E retains the full galleries and data-construction details.

4.2 Prediction across interaction domains (RQ2)

Figure 5 illustrates prediction differences at matched moments. In Deform360, **ASTER** keeps the separated object parts visible, whereas Cosmos 3 produces an elongated band. The MECKA potting-plants example compares hand–object configurations. Trainable models in this figure are fine-tuned on the same mixture with matched steps and GPU-hour budgets; the appendix retains all six models and 13 cases.

Table 1 summarizes the broader benchmark. **ASTER** has the lowest LPIPS and highest PSNR and SSIM in all eight settings. Average LPIPS reductions are 34.4% relative to Cosmos-Predict2.5 and 15.7% relative to Cosmos 3. The largest reductions over the backbone occur on GenieSim3.0 (47.7%) and YAM-XDof (30.3%), while the remaining domains also improve.

The advantage extends to DROID after target adaptation. With 10% target data, **ASTER** reaches 0.3016 LPIPS versus 0.3162 for Cosmos 3; with full target data, it reaches 0.2640 versus 0.2909. These results test transfer of the pretrained initialization, not zero-shot prediction. Appendix Table 2 reports all metrics, models, and uncertainty under the benchmark protocol.

4.3 Specified interactions and process validity (RQ3)

We adapt **ASTER** to painting uppercase letters and laser-cutting basic contours, then test seen targets, unseen shapes or sizes, and appearance changes. Here unseen means absent from task adaptation, not necessarily from foundation-model pretraining.

Inputs and evaluation. An offline code-generation agent writes a tool trajectory, which is checked in a geometry renderer and frozen before generation. Both models receive the same rendered, image-edited initial frame and task text; only **ASTER** receives point flow and camera geometry. This com-

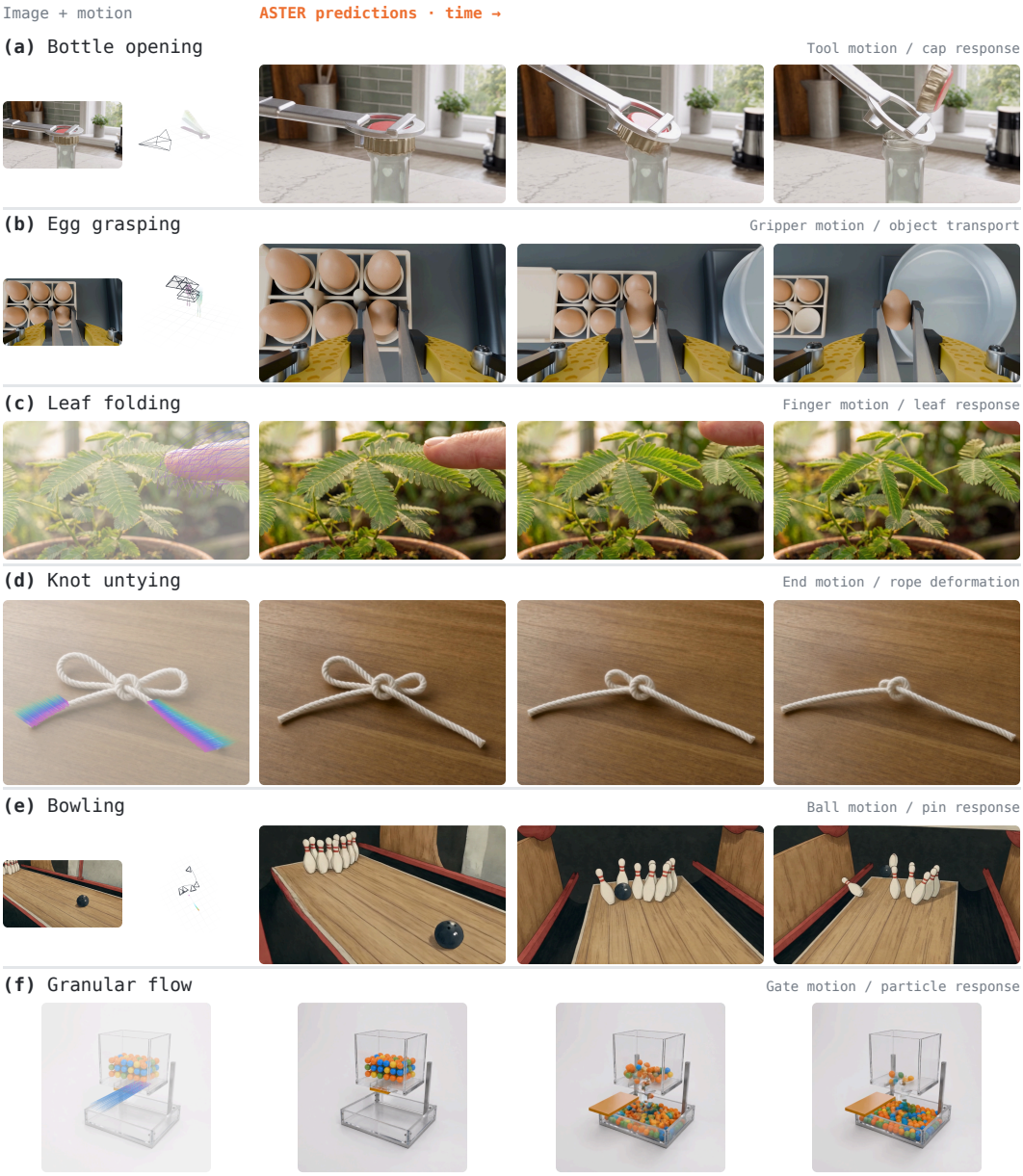

Figure 4: **Predicting the scene response to a moving body.** Six examples after task-focused post-training. Left: initial image and supplied motion, shown separately or with motion projected onto the image. Right: three ordered prediction frames from one source window per task. The affected cap, egg, leaves, rope, pins, and particles make the scene response visible. Full galleries appear in Appendix E.

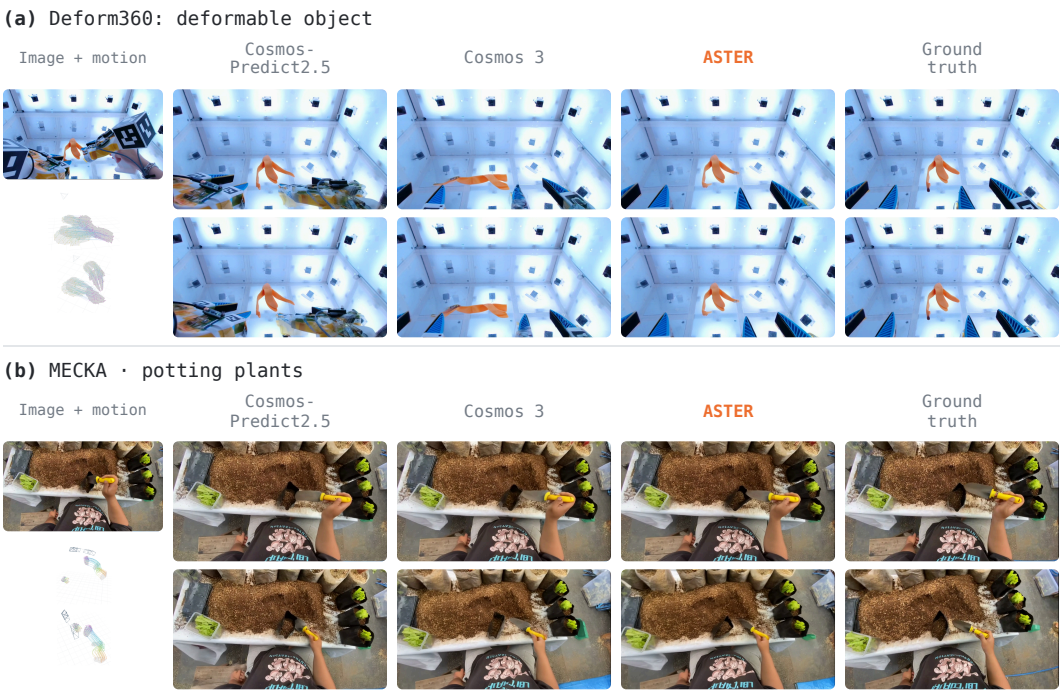

Figure 5: **Cross-domain video prediction.** Initial image and supplied geometry at left; two later moments for Cosmos-Predict2.5, Cosmos 3, **ASTER**, and ground truth. The examples expose differences in deformable-object shape and hand–object configuration. Trainable models here use the matched fine-tuning protocol of Appendix E.1; Table 1 uses its separate benchmark protocol. All models and 13 cases remain in the appendix.

pares task-specification interfaces rather than isolating pretraining. Two authors independently score all 48 videos per model with model identities hidden and presentation order randomized. Target accuracy checks final shape, position, and size; process validity checks whether marks form where and when the tool acts and persist without discontinuities. Agreement before adjudication is 90% for target accuracy and 88% for process validity.

Results. **ASTER** reaches the target in 40/48 cases (83%) versus 20/48 (42%) for Wan 2.6-I2V, and passes process validity in 35/48 (73%) versus 8/48 (17%). The distinction is important: 12 of the baseline's 20 target-correct videos fail the process criterion. On unseen targets, **ASTER** succeeds in 12/16 cases versus 4/16. Figure 6 illustrates the generated processes and appearance changes; Appendix Table 7 gives the full breakdown. These judgments measure visible execution in two controlled tasks, not material physics.

4.4 Compute and Data Composition (RQ4)

Compute. Five checkpoints span a 230× range of interaction-pretraining compute at fixed model size. All six source domains improve on LPIPS, PSNR, SSIM, and held-out rectified-flow loss at every evaluated checkpoint, without target adaptation (Appendix Figure 8). This is a training trajectory for one model size, not a scaling law across capacities.

Data composition. The full robot, human, and object mixture lowers LPIPS relative to Robot-only in all six domains, by up to 21.6%, including gains on robot domains. The recipes differ in training exposure as well as source composition, so the comparison does not isolate diversity at matched compute (Appendix C.5). After DROID adaptation, **ASTER** outperforms robot-only initialization with 10% or 100% target data, but robot-only is better with 1% (Appendix C.6).

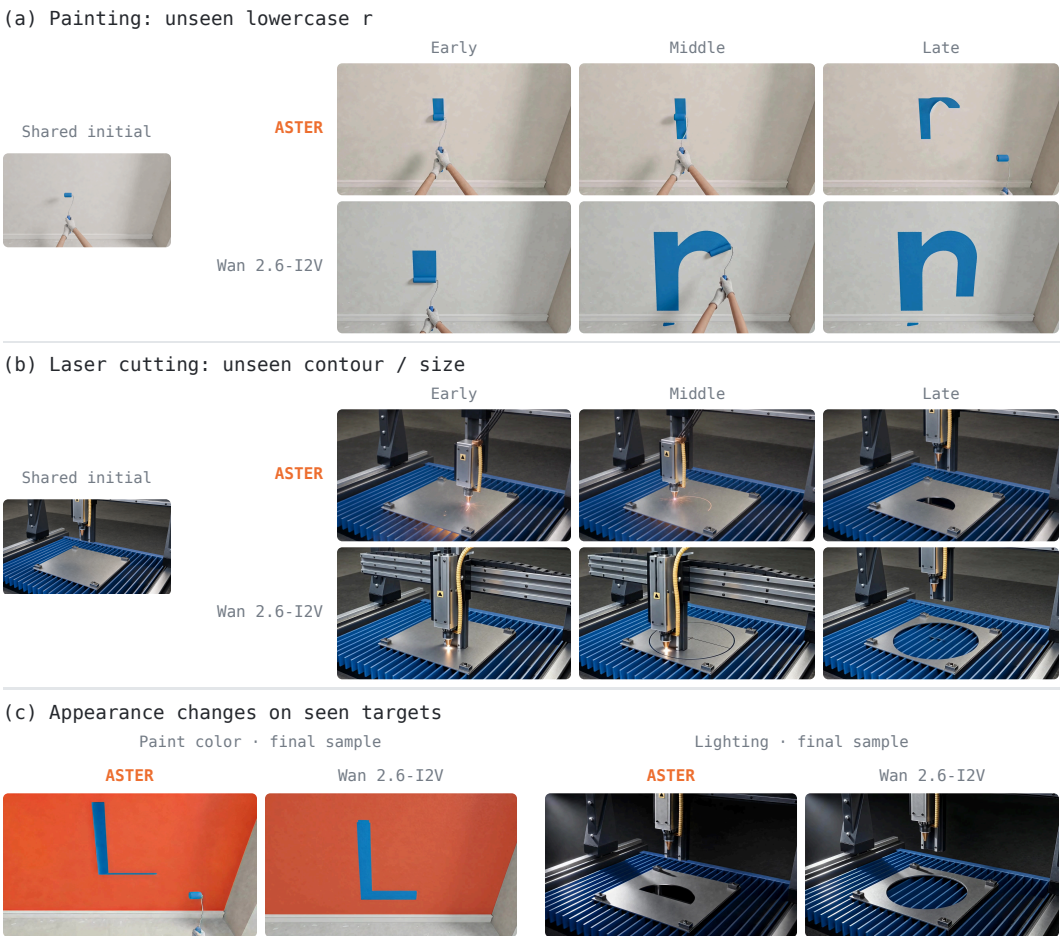

Figure 6: **Specified painting and laser-cutting interactions.** For each task, both models start from the same image and task text; **ASTER** additionally receives the point flow and camera. Process rows show unseen targets, while the bottom strip shows appearance changes on seen targets. The contrast highlights whether the generated marks follow the tool motion, not only whether the final frame resembles the requested shape.

Supporting studies. The initialization benefit persists with native robot actions: LPIPS falls from 0.3191 to 0.3114. Expanding adaptation beyond the restricted scope lowers LPIPS in 9 of 10 settings (Appendix C.7). Shuffling or removing motion also raises **ASTER**'s prediction error, particularly on YAM-XDof and GenieSim3.0, showing reliance on the supplied input. This diagnostic measures prediction of recorded futures, not the correctness of alternative outcomes (Appendix C.9).

5 Discussion and Limitations

A common motion-and-camera input lets heterogeneous experience train one video predictor. Reference-video metrics and process judgments provide complementary tests of its predictions, but neither establishes physical correctness.

Limitations. **ASTER** is not a validated physics simulator and has not been tested in closed-loop control. The showcases use selected post-training examples, and the execution study covers two tasks with human judgments. Scaling follows one run at one model size; external baselines have different inputs and training budgets. Robot-only pretraining remains better at the smallest DROID data fraction. Errors in tracking, depth, and camera estimation can also affect conditioning derived from general videos.

6 CONCLUSION

We presented **ASTER**, a video world simulator conditioned on mover point flow and camera geometry. This shared input allows joint pretraining on robot, human, and simulated interactions. Prediction improves across the evaluated domains, and task adaptation supports specified painting and cutting outcomes. The compute and data-composition studies show how additional interaction experience can improve a video predictor without requiring a common robot or native action space.

AI USE STATEMENT

Generative AI assisted literature discovery, brainstorming, study-design critique, manuscript editing, and LaTeX and analysis tooling. A vision-language model identified movers for general-video annotation (Appendix E.2). For the specified-interaction evaluation, an offline code-generation agent authored trajectory programs, which were checked in a geometry renderer, corrected as needed, and frozen before video generation; image editing restyled the rendered initial frames (Appendix D). Generated conceptual illustrations were used for visual planning. The authors remain responsible for the experiments, evidence, citations, and final claims.

ETHICS STATEMENT

This work studies interaction-conditioned video prediction for physical AI. Benefits include broader reuse of interaction experience; risks include misleading rollouts, privacy concerns in human recordings, and over-reliance on imperfect predictors in safety-critical settings. Generated video is not validated evidence of control safety. Dataset attribution and permitted use should accompany the exact releases and provider records used by the study.

REPRODUCIBILITY STATEMENT

Appendix A specifies the geometric representation, temporal alignment, conditioning, masks, loss, and optimization scopes. Appendix C provides evaluation definitions and supplementary results, and Appendix D describes the out-of-distribution study, including interaction authoring, initial-frame construction, and scoring. Numerical records and figure provenance are retained with the manuscript source.

REFERENCES

Alibaba Cloud. Alibaba Unveils Wan2.6 Series Enabling Everyone to Star in Videos, 2025. URL https://www.alibabacloud.com/en/press-room/alibaba-unveils-wan2-6-series-enabling-everyone. Press release, December 16; introduces Wan2.6-I2V.

Hadi Alzayer, Wenlong Huang, Haonan Chen, Christopher Luey, Lvmin Zhang, Maneesh Agrawala, Gordon Wetzstein, Li Fei-Fei, Yilun Du, Jiajun Wu, and Jia-Bin Huang. Masked Visual Actions for Unified World Modeling. *arXiv preprint arXiv:2607.19343*, 2026. URL https://arxiv.org/abs/2607.19343.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, et al. Genie 3: A New Frontier for World Models, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/. Google DeepMind, August 5.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Yixiang Chen, Peiyan Li, Jiabing Yang, Keji He, Xiangnan Wu, Yuan Xu, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. BridgeV2W: Bridging Video Generation Models to Embodied World Models via Embodiment Masks. *arXiv preprint arXiv:2602.03793*, 2026. URL https://arxiv.org/abs/2602.03793.

Kyle Chickering, Wei-An Lin, Swayam Bhanded, Dan Saunders, Akshat Tripathi, Jiaming Song, Shyamal Buch, and Xinchen Yan. Abra: Scaling Diffusion Image Training, 2026. URL https://arxiv.org/abs/2608.17286.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. Wan-Move: Motion-controllable Video Generation via Latent Trajectory Guidance. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=lHW93LKaUk.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi Fan. DreamDojo: A Generalist Robot World Model from Large-Scale Human Videos. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06949.

Chongjian Ge, Hanwen Jiang, Tianyu Wang, Jiuxiang Gu, Yiran Xu, Ziwen Chen, Shaoteng Liu, Jing Shi, Yicong Hong, Zefan Cai, Hailin Jin, and Hao Tan. Chimera: Designing and Chinchilla-Scaling Hybrid Visual Diffusion Transformers, 2026. URL https://arxiv.org/abs/2607.28611.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion Prompting: Controlling Video Generation with Motion Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.02700.

Bohai Gu, Taiyi Wu, Yueyang Yuan, Jian Liu, Xiaocheng Lu, Dazhao Du, Jie Zhang, Jinxiang Lai, Shuai Yang, Xiaotong Zhao, Alan Zhao, and Song Guo. WorldCraft: From Camera Navigation to Object Manipulation in Interactive Video World Models. *arXiv preprint arXiv:2605.25077*, 2026. URL https://arxiv.org/abs/2605.25077.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A Controllable Generative World Model for Robot Manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=748bHL2BAv.

David Ha and Jürgen Schmidhuber. World Models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Control Tasks through World Models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Adam W. Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Pavel Tokmakov, Suya You, Rares Ambrus, Katerina Fragkiadaki, and Leonidas J. Guibas. AllTracker: Efficient Dense Point Tracking at High Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://alltracker.github.io/.

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. URL https://openreview.net/forum?id=Z4evOUYrk7.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. URL https://arxiv.org/abs/2203.15556.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2601.03782.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking Generalization in Robot Learning through Video World Models. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5170–5194. PMLR, 2025. URL https://proceedings.mlr.press/v305/jang25a.html.

Zeren Jiang, Yushi Lan, Yihang Luo, Yufan Deng, Zihang Lai, Edgar Sucar, Christian Rupprecht, Iro Laina, Diane Larlus, Chuanxia Zheng, and Andrea Vedaldi. Syn4D: A Multiview Synthetic 4D Dataset. *arXiv preprint arXiv:2605.05207*, 2026. URL https://arxiv.org/abs/2605.05207.

Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10497–10509, 2025. doi: 10.1109/CVPR52734.2025.00982.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent Dynamic Depth from Stereo Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13229–13239, 2023. doi: 10.1109/CVPR52729.2023.01271.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://www.roboticsproceedings.org/rss20/p120.html.

Yejin Kim, Wilbert Pumacay, Omar Rayyan, Max Argus, Winson Han, Eli VanderBilt, et al. MolmoSpaces: A Large-Scale Open Ecosystem for Robot Navigation and Manipulation. *arXiv preprint arXiv:2602.11337*, 2026. URL https://arxiv.org/abs/2602.11337.

Klemen Kotar, Wanhee Lee, Rahul Venkatesh, Honglin Chen, Daniel Bear, Jared Watrous, Simon Kim, Khai Loong Aw, Lilian Naing Chen, Stefan Stojanov, Kevin Feigelis, Imran Thobani, Alex Durango, Khaled Jedoui, Atlas Kazemian, and Dan Yamins. World Modeling with Probabilistic Structure Integration. *arXiv preprint arXiv:2509.09737*, 2025. URL https://arxiv.org/abs/2509.09737.

Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. Generative Video Motion Editing with 3D Point Tracks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.02015.

Marion Lepert, Ria Doshi, and Jeannette Bohg. SHADOW: Leveraging Segmentation Masks for Cross-Embodiment Policy Transfer. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3536–3550. PMLR, 2025a. URL https://proceedings.mlr.press/v270/lepert25a.html.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training Robots Without Robots Using Only Human Videos. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4545–4565. PMLR, 2025b. URL https://proceedings.mlr.press/v305/lepert25a.html.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from In-the-wild Human Videos using Data-Editing. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2508.09976.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation. *arXiv preprint arXiv:2403.09227*, 2024. URL https://arxiv.org/abs/2403.09227.

Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2IV: Interaction-Centric Video Generation via Mask Trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a. URL https://arxiv.org/abs/2510.03135.

Hongyu Li, Wanjia Fu, Xiaoyan Cong, Zekun Li, Binghao Huang, Hanxiao Jiang, Xintong He, Yiqing Liang, Rao Fu, Tao Lu, Srinath Sridhar, Kevin A. Smith, George Konidaris, and Yunzhu Li. Deform360: A Massive Multi-view Visuotactile Dataset for Deformable World Models. In *European Conference on Computer Vision*, 2026b. URL https://arxiv.org/abs/2607.05390.

Hongyu Li, Bowen Wen, Xinghao Zhu, Yixuan Wang, Yilun Du, Yunzhu Li, George Konidaris, Stan Birchfield, Soha Pouya, Chenran Li, and Yan Chang. Hydra-0: Action Flow for Generalist World Modeling and Control. *arXiv preprint arXiv:2608.18077*, 2026c. URL https://arxiv.org/abs/2608.18077.

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling Laws for Diffusion Transformers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=T985gm4sDA.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do Generative Video Models Understand Physical Principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026. doi: 10.1109/WACV61042.2026.00099. URL https://arxiv.org/abs/2501.09038.

NVIDIA. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, et al. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025a. URL https://arxiv.org/abs/2501.03575.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, et al. World Simulation with Video Foundation Models for Physical AI. *arXiv preprint arXiv:2511.00062*, 2025b. URL https://arxiv.org/abs/2511.00062.

NVIDIA Cosmos Team. Cosmos-Predict2.5: Official Model and Inference Repository. https://github.com/nvidia-cosmos/cosmos-predict2.5, 2025.

SAM Team. SAM 3.1 Release Notes, 2026. URL https://github.com/facebookresearch/sam3/blob/main/RELEASE_SAM3p1.md. March 27 release; Object Multiplex video segmentation and tracking.

Tongxuan Tian, Haoyang Li, Bo Ai, Xiaodi Yuan, Zhiao Huang, and Hao Su. Diffusion Dynamics Models with Generative State Estimation for Cloth Manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 1703–1725. PMLR, 2025. URL https://proceedings.mlr.press/v305/tian25c.html.

Rahul Venkatesh, Honglin Chen, Kevin Feigelis, Daniel M. Bear, Khaled Jedoui, Klemen Kotar, Felix Binder, Wanhee Lee, Sherry Liu, Kevin A. Smith, Judith E. Fan, and Daniel L. K. Yamins. Understanding Physical Dynamics with Counterfactual World Modeling. In *Computer Vision – ECCV 2024*, pp. 368–387, 2024. doi: 10.1007/978-3-031-72691-0_21. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/3523_ECCV_2024_paper.php.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: Any Trajectory Instruction for Controllable Video Generation. *arXiv preprint arXiv:2505.22944*, 2025. URL https://arxiv.org/abs/2505.22944.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://vggt-omega.github.io/.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning Interactive Real-World Simulators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY.

Chenghao Yin, Da Huang, Di Yang, Jichao Wang, Nanshu Zhao, Chen Xu, Wenjun Sun, Linjie Hou, Zhijun Li, Junhui Wu, Zhaobo Liu, Zhen Xiao, Sheng Zhang, Lei Bao, Rui Feng, Zhenquan Pang, Jiayu Li, Qian Wang, and Maoqing Yao. Genie Sim 3.0: A High-Fidelity Comprehensive Simulation Platform for Humanoid Robot. *arXiv preprint arXiv:2601.02078*, 2026. URL https://arxiv.org/abs/2601.02078.

Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, Ke Lin, Jiarong Ou, Rui Chen, Victor Shea-Jay Huang, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, Wentao Zhang, and Kun Gai. Towards Precise Scaling Laws for Video Diffusion Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_Towards_Precise_Scaling_Laws_for_Video_Diffusion_Transformers_CVPR_2025_paper.html.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068. URL https://arxiv.org/abs/1801.03924.

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*, 2025a. doi: 10.15607/RSS.2025.XXI.015. URL https://arxiv.org/abs/2504.02792.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A Fine-Grained World Model for Robot Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9834–9844, 2025b. doi: 10.1109/ICCV51701.2025.00917. URL https://arxiv.org/abs/2406.14540.