SWoMo: Neuro-Symbolic World Model for Cataract Surgery Simulation

Ssharvien Kumar Sivakumar[1,2][0009−0009−0383−9462] Akwele Johnson[1], Anirudh Dhingra[1], Yannik Frisch[1,3], Ghazal Ghazaei[2], and Anirban Mukhopadhyay[1]

[1] Technical University Darmstadt, Darmstadt, Germany
ssharvien_kumar.sivakumar@tu-darmstadt.de
[2] Carl Zeiss AG, Munich, Germany
[3] AICM, Medical Faculty of Heidelberg University, Heidelberg University, Germany

Abstract. Realistic surgical simulation plays a crucial role in training novice surgeons and in the development of autonomous agents. World models can scale such simulation environments to realistic and diverse procedures by predicting future patient states conditioned on current observations and surgical actions. However, current state-of-the-art approaches often fail to satisfy key criteria required for clinical applicability, including visual realism, physically grounded interactions, and the ability to simulate scenarios beyond the training distribution. Hence, we introduce SWoMo, a neuro-symbolic world model for cataract surgery simulation that decouples motion generation from visual realism. The symbolic component, consisting of a rule-based simulator and scene graph representations, models motion dynamics and tool-tissue interactions, while a diffusion model produces realistic visual appearance, including textures and tissue deformations. We propose an inverse pairing strategy that reconstructs real surgical videos in the simulator to obtain paired simulated and real videos, which are then used to train our video diffusion model for the reverse objective of sim-to-real translation. Our experiments show both qualitative and quantitative improvements over prior work. We demonstrate that our simulator further satisfies the key criteria, including generalisation to unseen interaction geometries, improvements in downstream phase detection, and unsupervised video style transfer. The code, data, and model weights are available at: https://ssharvienkumar.github.io/SWoMo/.

Keywords: World Model · Video Generation · Interactive Simulation

1 Introduction

Surgery demands high precision, where irreversible actions must be carefully coordinated through structured spatial reasoning and causal understanding [6]. Simulated environments provide a safe and controllable setting in which such dependencies can be explored [8], enabling applications ranging from risk-free, immersive surgical skills training for novice surgeons to the development of autonomous surgical robotic agents that can reason and plan interactively [23,25].

However, scaling such environments to realistic and diverse procedures requires models that integrate continuous perception with explicit representations of the geometry and dynamics of the surgical scene [3]. Surgical world models address this need by enabling *interactive simulation* and *predictive modelling* of future patient states conditioned on current observations and actions [11,20], thereby *unifying perception, dynamics, and action* within a single framework [3].

However, the clinical applicability of a surgical world model hinges on satisfying three key criteria: **(I)** The simulated environment must achieve **high visual realism** to minimise the sim-to-real gap, as low fidelity limits policy transfer [18]. Yet this remains a major challenge for traditional simulators, particularly in modelling complex instrument-tissue interactions [20,28]. **(II)** Equally important is ensuring **physically grounded interactions**, so that both learning agents and human trainees acquire behaviours that transfer reliably to real high-stakes surgical settings [34]. **(III)** The **dynamic representation of the world model must go beyond the training distribution**, enabling the generation of plausible outcomes under unseen tool geometries, insertion angles, and novel surgical workflows.

Unfortunately, state-of-the-art research typically excels on one criterion while failing to satisfy others. Methods that attempt to control video generative models via various conditioning signals [2,12,26,28] fall short of interactive surgical simulation due to the lack of step-wise action conditioning. As a result, they provide limited fine-grained control for agent or trainee intervention and cannot be considered true world models. Their behaviour is also tightly coupled to the training distribution, degrading significantly under out-of-distribution (OOD) conditions [6]. World-model based approaches, such as SurgWM [20], inspired by Genie [4], learn interactive environments from unstructured surgical videos, but limit interaction to latent action codes rather than direct tool control. DreamGen [17] and SurgWorld [13] generate paired video-action data using inverse dynamics, improving policy learning, yet they lack physical grounding. Unpaired image translation methods [24,32] have also been explored to address the sim-to-real gap, but remain insufficient for modelling complex spatiotemporal tool-tissue interactions.

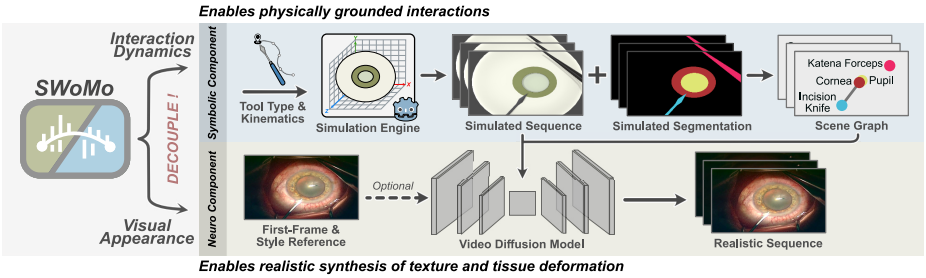

Fig. 1. Neuro-symbolic World Model for interactive cataract surgery simulation that decouples surgical interaction dynamics from visual appearance.

Hence, we introduce SWoMo, an interactive neuro-symbolic world model for cataract surgery, as shown in Figure 1, in which **symbolic scene graphs and a rule-based physics simulator** explicitly encode structure and constraints, while a **diffusion model** provides high-fidelity visual synthesis. Our proposed method *explicitly decouples the modelling of surgical motion and interaction dynamics from visual appearance*. This allows precise tool and anatomical motions to be preserved through physically grounded simulation while enabling the generation of entirely novel scenarios beyond the training distribution. Visual realism, such as texture and tissue deformation, is learned through data-driven synthesis using a diffusion model. To achieve this, we introduce **inverse pairing strategy** in which tool and anatomical motions are extracted from real surgical videos and replayed in the simulator, producing large-scale paired simulated and real videos. These pairs are then used to train a video-to-video diffusion model that translates simulated renderings back into realistic surgical videos. The simulator also produces corresponding segmentations, which are used to construct scene graphs that efficiently encode the scene and object relationships in a structured graphical format. We show that conditioning the diffusion model on scene graphs is crucial for mitigating issues caused by residual misalignment between simulated and real videos. Optional initial frame or style reference conditioning is added for subject-specific surgical simulation. We further demonstrate generalisation to unseen surgical interactions, notable improvements in downstream phase detection, and unsupervised video style transfer across datasets.

2 Method

Our surgical world model decouples motion generation from visual realism. Motion and interaction dynamics are governed by symbolic components comprising of scene graphs $\mathcal{G}$ and a rule-based simulator $T(\cdot)$, while visual appearance is generated by a denoising diffusion model, employing a neural network $\boldsymbol{\epsilon}_\theta(\cdot)$. We begin by extracting the anatomical configuration k_t^{anat} and tool kinematics k_t^{tool} from real surgical video $x_{1:T}$, and use them to construct a digital twin of the eye and replay these signals in this simulator. At time step t, we define the simulator state as $\bar{x}_t = \{k_t^{\text{anat}}, k_t^{\text{tool}}\}$. Actions correspond to explicit surgical tool motions, $a_t = \Delta k_t^{\text{tool}}$. State transitions are governed by $T(\cdot)$ such that $(\bar{x}_{t+1}, \bar{m}_{t+1}) = T(\bar{x}_t, a_t)$, where $\bar{x}_{t+1}$ denotes the next state along with its rendered simulated frame, and $\bar{m}_{t+1}$ is the corresponding segmentation. From each simulator state, a scene graph is generated, $\mathcal{G}_t = S(\bar{x}_t, \bar{m}_t)$, encoding object geometry and relational structure. Final observations are then generated according to $p_\theta(x_{t:t+n} \mid \bar{x}_{t:t+n}, \mathcal{G}_{t:t+n})$, parameterized by the diffusion model, which translates simulated sequence $\bar{x}_{t:t+n}$ into realistic sequence $x_{t:t+n}$. Here, n denotes the number of consecutive simulated frames jointly provided as conditioning to the diffusion model.

Inverse Pairing: In this section, we describe the extraction of k_t^{anat} and k_t^{tool} from real surgical video $x_{1:T}$. We start by segmenting pupil and iris employing nnU-Net [16], and corresponding segmentations are fitted with ellipses to

obtain compact geometric parameters k_t^{iris} and k_t^{pupil} encoding centroid, orientation, and axis lengths. Motion of the eye globe decomposes into global and local components. *Global motion*, arising from camera or patient movement, is estimated by annotating skin landmarks in the first frame x_1 and tracking them over time using pre-trained CoTracker [19]. The resulting trajectories define a global transformation g_t used to estimate the displacement of the eye globe centroid. *Local motion* corresponds to rotational movement of the eye globe and is derived from the centroid trajectory of the pupil mask, yielding rotational parameters r_t. Formally, the globe motion is represented as $k_t^{\text{globe}} = \{g_t, r_t\}$. The anatomical configuration is thus defined as $k_t^{\text{anat}} = \{k_t^{\text{globe}}, k_t^{\text{iris}}, k_t^{\text{pupil}}\}$. Tool kinematics are recovered from tool masks generated using SASVi [29] and SAM2-based [27] manual interactive annotation tool. From these masks, we extract geometric parameters $k_t^{\text{tool}} = \{c_t, \theta_t, \beta_t\}$ where c_t denotes tool tip position, θ_t orientation, and β_t articulation parameters such as bending angle and opening angle.

Rule-based Simulator: The recovered anatomical and tool parameters $\{k_t^{\text{anat}}, k_t^{\text{tool}}\}$ are used to drive a rule-based simulator $T(\cdot)$ on Godot game engine [10]. The simulator instantiates a parameterised eye globe and surgical tools, whose mesh transformations are directly controlled by these parameters. The global translation g_t is mapped from image coordinates to simulator space through normalisation and fixed scaling, defining the eye globe position, while rotational parameters r_t control eye globe orientation via yaw and pitch. Anatomical meshes are updated using k_t^{iris} and k_t^{pupil}. Iris and pupil geometry are scaled according to their estimated axis lengths, enabling subject-specific variation.

Tool meshes are controlled by $k_t^{\text{tool}} = \{c_t, \theta_t, \beta_t\}$. Tool tip positions c_t are mapped to simulator space using the same normalisation as global globe motion, while orientations θ_t directly determine tool rotation. For articulated instruments such as forceps, articulation parameters β_t control relative mesh rotations to reproduce opening motions, while for angled tools they adjust the shaft bend to maintain geometric consistency. Final tool positions are defined relative to the eye globe surface with an offset scaled by the globe's anatomical scaling factor, ensuring consistent placement in the simulated video $\bar{x}_{1:T}$. We present a real video and its simulated pair, along with segmentations, in Supplementary E.

Symbolic Scene Graph: We form scene graphs $\mathcal{G}_{1:n}$ from $\bar{x}_{1:n}$ and its segmentation $\bar{m}_{1:n}$ and subsequently encode $\mathcal{G}_{1:n}$ following a strategy inspired by SG2VID [28]. Each node in $\mathcal{G}_{1:n}$ represents a connected component from $\bar{m}_{1:n}$ and stores high-level component attributes, including centroid, spatial spreading, and average optical flow within the corresponding region. We pre-trained two separate graph encoders, global encoder $E_{\mathcal{G}}^{glob}$ and local encoder $E_{\mathcal{G}}^{loc}$. The $E_{\mathcal{G}}^{glob}$ learns high-level structural relationships via contrastive learning between $\mathcal{G}_{1:n}$ and $\bar{m}_{1:n}$. In contrast, $E_{\mathcal{G}}^{loc}$ focuses on appearance cues by learning to reconstruct masked regions in the real sequence $x_{1:n}$ using information from $\mathcal{G}_{1:n}$. We find the additional $\mathcal{G}_{1:n}$ conditioning to be important for visual fidelity and conditional adherence, as also highlighted by the ablation study in Table 1. It contributes in two main ways: *first*, $\mathcal{G}_{1:n}$ provides explicit class-level information that is not available from the $\bar{x}_{1:n}$ alone, helping the model distinguish object

identities. *Second*, we observe regions near component boundaries, especially near tool-tissue interaction, often suffer from degraded synthesis due to misalignments between $\bar{x}_{1:n}$ used for video conditioning and $x_{1:n}$, which introduce conflicting training signals. When such inconsistencies occur repeatedly, the model tends to average visual features across boundaries, causing smoothing artifacts. Incorporating $\mathcal{G}_{1:n}$ alleviates this by abstracting local pixel-level misalignments, as the node's component-level attributes remain relatively stable under small boundary shifts. Moreover, $\mathcal{G}_{1:n}$ provides only higher-level structural guidance, allowing generator to infer realistic visual boundaries by itself from the learned appearance priors rather than strictly following misaligned $\bar{x}_{1:n}$ conditioning.

Video Diffusion Model with Graph-Image-Video Conditioning: Our diffusion model for translating simulated sequence $\bar{x}_{t:t+n}$ into realistic sequence $x_{t:t+n}$ is trained in two stages, as illustrated in Figure 2: *first*, learning image and graph to video generation, and *subsequently* incorporating video-level conditioning through ControlNet [35] training. Diffusion models [14] rely on a parameterised network $\boldsymbol{\epsilon}_\theta$ that is trained to reverse the gradual noise injection process $p(x_{t,\tau-1} \mid x_{t,\tau}, c)$, where τ denotes the diffusion timestep. Training is performed by minimizing a mean squared error objective between the true noise and the noise predicted by $\boldsymbol{\epsilon}_\theta$: $\min_\theta \mathbb{E}_{\tau,x_{t,0},\epsilon}\left[\|\epsilon - \epsilon_\theta(x_{t,\tau}, \tau, c)\|^2\right]$. To extend diffusion models to video generation [15], the model is augmented with temporal layers. Specifically, temporal convolution and attention layers are interleaved with spatial layers, enabling the model to capture temporal dependencies and motion dynamics across frames. In the first stage, the conditioning signal c consists of the first-frame x_1 and the sequence scene graphs $\mathcal{G}_{1:n}$. For first-frame conditioning, the noise term ϵ_1 is replaced with the actual first-frame x_1. The resulting model input is therefore constructed as $\hat{\boldsymbol{\epsilon}} = \{x_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n\}$. For graph conditioning, we concatenate the outputs of $E_{\mathcal{G}}^{glob}$ and $E_{\mathcal{G}}^{loc}$ and further concatenate the result with the timestep embedding before passing it to $\boldsymbol{\epsilon}_\theta$.

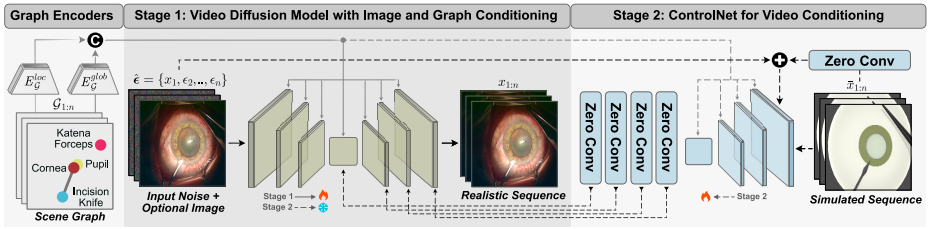

Fig. 2. Overview of SWoMo's two-stage video diffusion training.

In the *second stage*, we enable conditioning on the simulated sequence $\bar{x}_{1:n}$. For that, we freeze the parameters θ of the pre-trained diffusion backbone $\boldsymbol{\epsilon}_\theta$ from the previous stage and create a separate, trainable copy of its encoder with parameters θ_c. The frozen backbone $\boldsymbol{\epsilon}_\theta$ and the trainable encoder are connected through zero-initialised convolutional layers at multiple resolutions [35], allow-

ing the conditioning simulated sequence $\bar{x}_{1:n}$ to modulate intermediate feature representations without disrupting the pretrained generative prior of $\boldsymbol{\epsilon}_\theta$ and preserving its visual quality. The $\bar{x}_{1:n}$ is encoded through this control branch, and the resulting features are injected into the corresponding layers of $\boldsymbol{\epsilon}_\theta$.

3 Experiments and Results

We evaluate our method on two publicly available cataract surgery datasets: CATARACTS [1] and Cataract-1k [9]. To limit the need for extensive frame-level annotations, we restrict our modelling to safety-critical phases, which together account for around half of the video: *Idle, Incision, Viscoelastic, Capsulorhexis, Hydrodissection*, and *Phacoemulsification*. After this filtering, the datasets contain 868 videos from Cataract-1k and 50 videos from CATARACTS, with a mean duration of just over three minutes. For training, videos are temporally sampled at 4 frames per second into sequences of 16 frames and spatially resized to 128×128. Data splits are created at the video level, using a 50/6/44 ratio for CATARACTS and 80/10/10 for Cataract-1k. The training of the diffusion model is distributed across four NVIDIA A40 GPUs.

Table 1. Quantitative Comparisons of Synthesis Quality and Conditioning Adherence.

	CATARACTS [1]					**Cataract-1k** [9]				
Method	FVD↓	FID↓	LPIPS↑	BB IoU↑	F1↑	FVD↓	FID↓	LPIPS↑	BB IoU↑	F1↑
StyleGAN-V [30]	581.5	107.2	0.379	–	–	501.2	116.4	0.286	–	–
Endora [21]	436.8	58.4	0.456	–	–	258.7	40.0	0.379	–	–
MedSora [33]	1243.3	127.6	0.403	–	–	809.9	147.6	0.324	–	–
LVDM [12]	1604.6	131.0	0.557	0.228	0.154	1469.6	176.6	0.519	0.213	0.188
MOFA [26]	993.4	105.6	0.446	0.432	0.282	716.7	94.6	0.358	0.418	0.404
SG2VID [28]	363.8	47.3	0.436	0.497	0.391	73.0	14.9	0.392	0.607	0.623
SWoMo	265.4	40.8	0.450	0.522	0.412	123.0	20.1	0.388	0.645	0.656
SWoMo-✗IMG	329.3	42.3	0.451	0.514	0.406	134.3	20.2	0.389	0.622	0.642
SWoMo-✗SG	390.7	52.1	0.463	0.377	0.296	283.6	35.5	0.434	0.529	0.554

Quantitative Comparison and Ablation: We evaluate the quality and diversity of synthesised videos using FID, FVD [31], and the LPIPS diversity score [36], as shown in Table 1. For conditional methods, we additionally measure adherence to conditioning using detection-based metrics, including F1 score and bounding box IoU. Specifically, we train Mask2Former [7] to detect tools and anatomical structures, and compare bounding box predictions across synthesised and real videos. Our method is benchmarked against multiple baselines with publicly available implementations. StyleGAN-V [30], Endora [21], and MedSora [33] represent unconditional video generation approaches. We further modify LVDM [12] to enable text-based conditioning using scene graph triplets, whereas SG2VID [28] directly conditions on scene graphs. MOFA [26] conditions generation on both the initial frame and motion trajectory derived

from sparse optical flow. SWoMo outperforms most baselines in terms of visual fidelity and achieves substantial improvements over all baselines in terms of conditioning adherence. In Table 1, we also provide ablations without scene graph conditioning (SWoMo-✗SG) and without image conditioning (SWoMo-✗IMG) to illustrate the contribution of each conditioning component.

Qualitative Assessment: We present SWoMo's qualitative results in Figure 3, with additional samples in Supplementary A and a qualitative comparison in Supplementary B. SWoMo leverages both the simulated video and an initial frame for video synthesis, whereas SWoMo-✗IMG relies solely on the simulated video. These results highlight how SWoMo's sim-to-real transfer task reduces major implementation effort on the simulator, for example, by eliminating the need to explicitly model textures and deformation, while still accurately following tool movements and anatomical configurations from the simulated video.

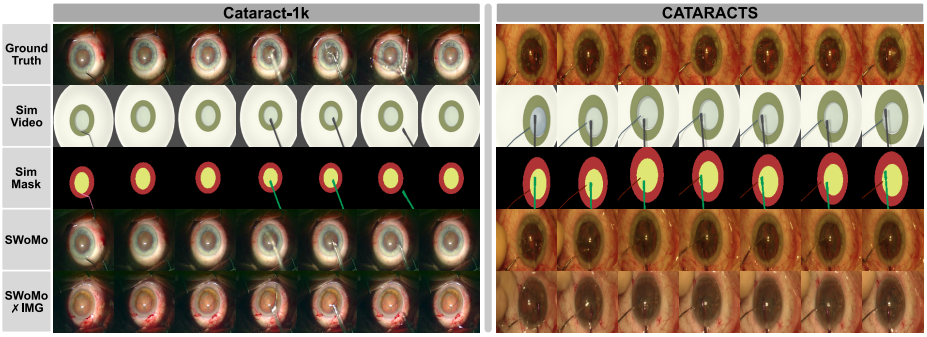

Fig. 3. Qualitative Results of Sim-to-Real Video Transfer.

Unsupervised Video Style Transfer: SWoMo enables unsupervised style transfer between the CATARACTS and Cataract-1k domains without training on paired sequences. This is achieved through a shared intermediate representation of the simulated sequence that is used across both datasets. Specifically, we utilise the simulated sequence from the source domain with the initial-frame conditioning from the target domain to synthesise videos in the target style. Because the intermediate representation encodes geometry and motion independently of appearance, the model preserves interaction dynamics while adopting the visual characteristics of the target domain. The style transfer results are shown in Figure 4 (top) for both datasets, with full videos provided in Supplementary C.

Improved Generalisation to Novel Tool Motions: Generative models are fundamentally constrained by their training distributions and often degrade on sequences far outside these distributions. By introducing an intermediate simulated sequence representation that provides explicit structural and motion guidance, we further push the boundaries of what the model can handle. Using the simulator, we generate difficult cases, including tools with novel entry directions and tool combinations that never co-occur in the training data. Visual

results are shown in Figure 4 (bottom), with additional examples provided in Supplementary D, demonstrating that the intermediate representation substantially improves generalisation under OOD conditions.

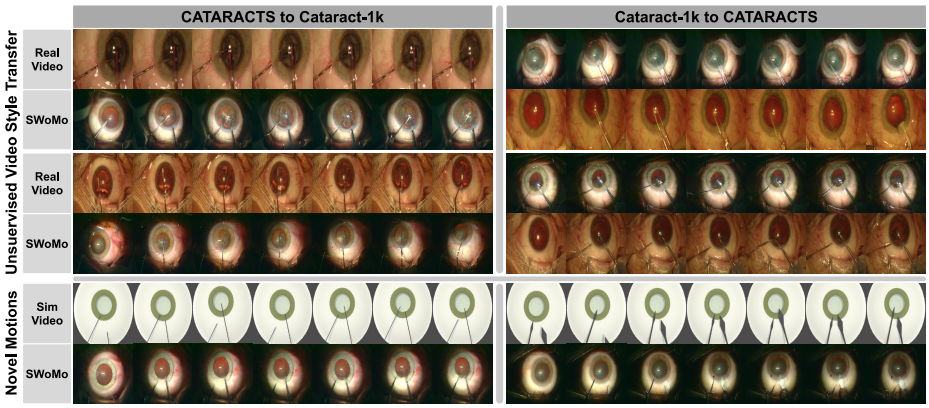

Fig. 4. Unsupervised Video Style Transfer and Generalisation to Novel Tool Motion

Downstream Evaluation on Phase Recognition: We use synthesised videos to augment the training data of a downstream model for phase recognition during cataract surgery. The phase recognition is performed using MS-TCN++ [22] trained on DINO features [5]. To generate a synthesised video, two real videos are randomly selected from the training set. The visual style is taken from the first video, while the tool and anatomical motion patterns are taken from the second. Phase annotations are inherited from the second video, since the underlying surgical actions remain unchanged, making this process a form of generative data augmentation. To generate full-length surgical videos, sequences are synthesised autoregressively by first generating a sequence and then using its last frame as the first-frame conditioning input for the subsequent sequence. Using this strategy, we effectively double the number of training videos. Table 2 presents the results, demonstrating improvements over training on real data alone and also over generative augmentation with other methods.

Table 2. Performance on Downstream Phase Recognition.

	CATARACTS [1]		**Cataract-1k** [9]	
Training data	Accuracy↑	F1-Score↑	Accuracy↑	F1-Score↑
Real Only	79.4	79.3	93.7	94.9
Real + LVDM [12]	63.6 (-15.8)	65.3 (-14.0)	71.3 (-22.4)	72.4 (–22.5)
Real + SG2VID [28]	80.5 (+1.1)	81.6 (+2.3)	94.0 (+0.3)	95.3 (+0.4)
Real + SWoMo (Ours)	82.5 (+3.1)	83.0 (+3.7)	94.2 (+0.5)	95.2 (+0.3)

4 Conclusion

We present SWoMo, a neuro-symbolic world model for cataract surgery in which tool kinematics are fed into a rule-based simulator, whose outputs are then converted into realistic videos, effectively combining the strengths of physical simulation and generative modelling. SWoMo fulfils the role of a world model; given the current simulation state and action input, it predicts the next state and generates the corresponding future video frame. Unlike world models that rely primarily on implicit or latent representations, SWoMo explicitly models scene dynamics, which is particularly important for physically grounded interactions and improved generalisation beyond the training distribution, where purely implicit world models may struggle. We demonstrate several unique capabilities of SWoMo, including precise unsupervised video style transfer without additional training and strong generalisation to unseen interaction geometries with novel tool motions and co-occurrences. Finally, we show that the generated videos can effectively augment existing datasets and improve performance on downstream tasks such as surgical phase recognition.

Acknowledgments. The work is partially funded by the German Federal Ministry of Education and Research as part of the Software Campus (research grant 01|S23067).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., et al.: Cataracts: Challenge on automatic tool annotation for cataract surgery. MedIA **52**, 24–41 (2019)
2. Biagini, D., Navab, N., Farshad, A.: Hierasurg: Hierarchy-aware diffusion model for surgical video generation. In: MICCAI. pp. 310–319. Springer (2025)
3. Boels, M., Robertshaw, H., Booth, T.C., Granados, A., Dasgupta, P., Ourselin, S.: Surgical robot learning: From demonstration and simulation to world models-a review. Authorea Preprints (2025)
4. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: ICML (2024)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
6. Chen, Z., Xu, Q., Wu, J., Yang, B., Zhai, Y., Guo, G., Zhang, J., Ding, Y., Navab, N., Luo, J.: How far are surgeons from surgical world models? a pilot study on zero-shot surgical video generation with expert assessment. arXiv:2511.01775 (2025)
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1280–1289 (2021)

8. Frisch, Y., Sivakumar, S.K., Köksal, Ç., Böhm, E., Wagner, F., Gericke, A., Ghazaei, G., Mukhopadhyay, A.: Surgrid: controllable surgical simulation via scene graph to image diffusion. Int J CARS **20**(7), 1421–1429 (2025)
9. Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R.: Cataract-1k: cataract surgery dataset for scene segmentation, phase recognition, and irregularity detection. arXiv:2312.06295 (2023)
10. Godot Engine Contributors: Godot engine (2024), `https://godotengine.org`, free and open-source 2D and 3D game engine
11. Ha, D., Schmidhuber, J.: World models. arXiv:1803.10122 **2**(3) (2018)
12. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv:2211.13221 (2022)
13. He, Y., Guo, P., Xu, M., Li, Z., Myronenko, A., Imans, D., Liu, B., Yang, D., Gu, M., Ji, Y., et al.: Surgworld: Learning surgical robot policies from videos via world modeling. arXiv:2512.23162 (2025)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. NeurIPS **35**, 8633–8646 (2022)
16. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
17. Jang, J., Ye, S., Lin, Z., Xiang, J., Bjorck, J., Fang, Y., Hu, F., Huang, S., Kundalia, K., Lin, Y.C., et al.: Dreamgen: Unlocking generalization in robot learning through video world models. arXiv:2505.12705 (2025)
18. Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., Batra, D.: Sim2real predictivity: Does evaluation in simulation predict real-world performance? IEEE Robot Autom. Let. **5**(4), 6670–6677 (2020)
19. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is 472 better to track together. arXiv:2307.07635 **473** (2023)
20. Koju, S., Bastola, S., Shrestha, P., Amgain, S., Shrestha, Y.R., Poudel, R.P., Bhattarai, B.: Surgical vision world model. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 1–10. Springer (2025)
21. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. In: MICCAI. pp. 230–240. Springer (2024)
22. Li, S., Farha, Y.A., Liu, Y., Cheng, M.M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE TPAMI (2020)
23. Lin, H., Li, B., Au, K.W.S.: Visuomotor grasping with world models for surgical robots. arXiv:2508.11200 (2025)
24. Martyniak, S., Kaleta, J., Dall'Alba, D., Naskręt, M., Płotka, S., Korzeniowski, P.: Simuscope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. In: WACV. pp. 4268–4278. IEEE (2025)
25. Nair, A.G., Ahiwalay, C., Bacchav, A.E., Sheth, T., Lansingh, V.C., Vedula, S.S., Bhatt, V., Reddy, J.C., Vadavalli, P.K., Praveen, S., et al.: Effectiveness of simulation-based training for manual small incision cataract surgery among novice surgeons: a randomized controlled trial. Scientific reports **11**(1), 10945 (2021)
26. Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In: ECCV. pp. 111–128. Springer (2024)

27. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv:2408.00714 (2024)
28. Sivakumar, S.K., Frisch, Y., Ghazaei, G., Mukhopadhyay, A.: Sg2vid: Scene graphs enable fine-grained control for video synthesis. In: MICCAI. pp. 511–521. Springer (2025)
29. Sivakumar, S.K., Frisch, Y., Ranem, A., Mukhopadhyay, A.: Sasvi: segment any surgical video. Int J CARS **20**(7), 1409–1419 (2025)
30. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: CVPR. pp. 3626–3636 (2022)
31. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717 (2018)
32. Venkatesh, D.K., Rivoir, D., Pfeiffer, M., Speidel, S.: Surgical-cd: Generating surgical images via unpaired image translation with latent consistency diffusion models. In: European Conference on Computer Vision. pp. 218–235. Springer (2024)
33. Wang, Z., Zhang, L., Wang, L., Zhu, M., Zhang, Z.: Optical flow representation alignment mamba diffusion model for medical video generation. arXiv:2411.01647 (2024)
34. Yang, Y., Zhang, Z., Zhang, X., Zeng, Y., Li, H., Zuo, W.: Physworld: From real videos to world models of deformable objects via physics-aware demonstration synthesis. arXiv:2510.21447 (2025)
35. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)