SAW: Toward a Surgical Action World Model via Controllable and Scalable Video Generation

Sampath Rapuri[1,⋆], Lalithkumar Seenivasan[1,⋆], Dominik Schneider[1], Roger Soberanis-Mukul[1], Yufan He[2], Hao Ding[1], Jiru Xu[1], Chenhao Yu[1], Chenyan Jing[1], Pengfei Guo[2], Daguang Xu[2], and Mathias Unberath[1]

[1] Johns Hopkins University, Baltimore MD, USA
[2] NVIDIA, Santa Clara, USA

Abstract. A surgical world model capable of generating realistic surgical action videos with precise control over tool-tissue interactions can address fundamental challenges in surgical AI and simulation – from data scarcity and rare event synthesis to bridging the sim-to-real gap for surgical automation. However, current video generation methods, the core of such surgical world models, require expensive annotations or complex structured intermediates as conditioning signals at inference, limiting their scalability. Other approaches exhibit limited temporal consistency across complex laparoscopic scenes and do not possess sufficient realism. We propose Surgical Action World (SAW) – a step toward surgical action world modeling through video diffusion conditioned on four lightweight signals: language prompts encoding tool-action context, a reference surgical scene, tissue affordance mask, and 2D tool-tip trajectories. We design a conditional video diffusion approach that reformulates video-to-video diffusion into trajectory-conditioned surgical action synthesis. The backbone diffusion model is fine-tuned on a custom-curated dataset of 12,044 laparoscopic clips with lightweight spatiotemporal conditioning signals, leveraging a depth consistency loss to enforce geometric plausibility without requiring depth at inference. SAW achieves state-of-the-art temporal consistency (CD-FVD: 199.19 vs. 546.82) and strong visual quality on held-out test data. Furthermore, we demonstrate its downstream utility for (a) surgical AI, where augmenting rare actions with SAW-generated videos improves action recognition (clipping F1-score: 20.93% to 43.14%; cutting: 0.00% to 8.33%) on real test data, and (b) surgical simulation, where rendering tool-tissue interaction videos from simulator-derived trajectory points toward a visually faithful simulation engine.

Keywords: Surgical World Model · Surgical Video Generation · Deep Learning · Action Recognition · Surgical Simulator.

1 Introduction

A surgical world model that enables controllable, scalable video synthesis of surgical actions with realistic tool-tissue interaction has the potential to transform

⋆ Equal contribution.

both surgical AI and the next generation of high-fidelity surgical simulators. For surgical AI, it offers a path to overcome the fundamental data scarcity that is a bottleneck for developing perception models, and enables targeted synthesis of rare but clinically critical events without costly setups or tissue use [26, 33, 24, 20]. For simulation, such a generative world model can bridge the persistent sim-to-real gap by enabling visually realistic closed-loop environments where instrument motion drives controllable tool-tissue interactions. This supports both accelerated curation of paired kinematics-video data necessary for automating surgical actions and the development of digital twins for surgical safety evaluation through the forward simulation of tool actions [22, 19, 10, 30]. Realizing such world models hinges on advances in controllable and realistic video generation. While recent surgical video generation methods [2, 21, 6] have demonstrated potential in producing visually plausible outputs, they remain limited in inference-time controllability and scalability: HieraSurg relies on expensive dense annotations such as per-frame segmentation masks [2]; SG2VID depends on structured intermediates (spatio-temporal scene graphs) that are difficult to obtain or manipulate at inference [21]; SurgVisionWM does not allow for interpretable, user-defined action conditioning signals [13]; and SurgSora, which employs flexible trajectory-conditioning, is constrained by limited inference windows ($W = 21$) and struggle to maintain temporal consistency [6]. Reinforcement Learning-based surgical world models that learn dynamics in low-dimensional state spaces have also been explored [8], but differ fundamentally from generative world models that treat video as state and condition on user-defined actions.

To this end, we propose Surgical Action World (SAW), a step toward a surgical action world model aligned with the modern generative formulation [3, 4]. SAW enables controllable and scalable video diffusion conditioned on four lightweight signals: (i) structured language prompts encoding instrument and action context, (ii) a reference first frame anchoring scene appearance, (iii) tissue affordance masks specifying interaction regions, and (iv) 2D tool-tip trajectory across the surgical scene. **Our key contributions are**: (1) a curated dataset of 12,044 laparoscopic video clips with spatiotemporal annotations including video-level tool-action labels and tissue affordance along with frame-level tool-tip trajectories; (2) a video diffusion approach that reformulates video-to-video generation for scalable, highly controllable surgical action synthesis without relying on expensive inference-time conditioning signals, and incorporates a novel depth consistency loss enforcing geometric plausibility without requiring depth at inference; and (3) demonstration of two downstream applications – augmentation of rare surgical actions to improve action recognition on real test data, and rendering realistic tool-tissue interactions from simulator-derived tool trajectories to drive visually faithful surgical simulation.

2 Surgical Action World (SAW) Model

To advance surgical world modeling, we propose Surgical Action World (SAW), which reformulates video-to-video diffusion for controllable surgical action syn-

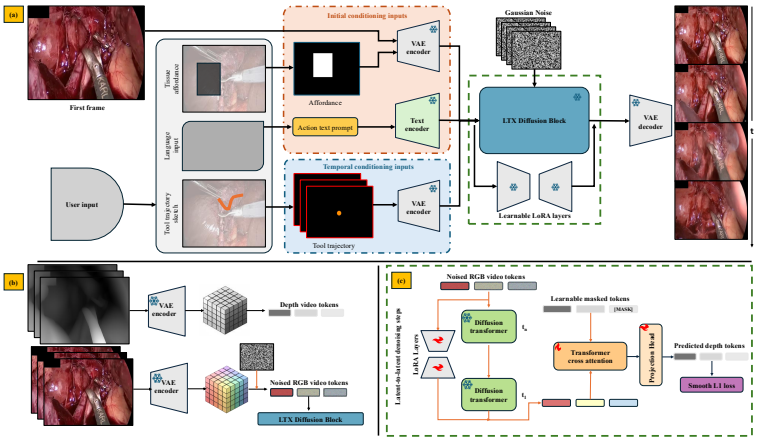

Fig. 1: **a)** At inference, SAW leverages a user-inputted 2D tool trajectory, language prompt, tissue affordance, and a background surgical scene, enabling realistic and controllable video diffusion. **b)** During training, normalized depth and RGB videos are encoded using a frozen VAE, while the RGB video embedding is then flattened and noised into a sequence of noisy tokens. The LoRA layers within the diffusion transformer are updated. **c)** During training, we compute a depth consistency loss ($\mathcal{L}_{\text{DC}}$) as a Smooth L1 loss between predicted masked depth video tokens and pseudo-ground-truth depth tokens derived solely from the denoised RGB video tokens.

thesis. By conditioning on instrument trajectories and tissue affordance alongside standard text and image inputs, SAW models tool-tissue interaction dynamics for scalable and realistic video generation (Fig. 1). The diffusion process is defined as $z_{i-1} = \mathcal{D}(z_i, t_i, \mathbf{z^a}, \mathbf{z^f}, \mathbf{z^p}, \mathbf{z}^\gamma)$, where z_i is the latent embedding from the previous denoising stage, t_i the denoising iteration with $i = 1, \ldots, N = 50$, and $z_0 = x_0$ the clean video. Generation is conditioned on four lightweight signals: a language prompt $\mathbf{z^a}$ encoding tool-action context, a reference frame $\mathbf{z^f}$ anchoring scene appearance, an instrument trajectory $\mathbf{z^p}$ controlling tool-tip motion, and a tissue affordance map $\mathbf{z}^\gamma$ specifying interaction regions, each of which can be provided independently or in combination at inference.

2.1 Video Diffusion Model: We adopt LTX-Video [11] as the backbone for SAW, a transformer-based latent diffusion model that natively supports multimodal conditioning (text, video, image). LTX performs diffusion in latent space using a variational autoencoder (VAE) for spatiotemporal downsampling, with denoising via a diffusion transformer trained using flow-matching [15]. We fine-tune LTX using In-Context Low Rank Adaptation (IC-LoRA) [12], a parameter-efficient fine-tuning strategy to fine-tune the model via learnable low-rank weight adaptations, for surgical action synthesis with four conditioning embeddings: a reference frame $\mathbf{z^f}$, language prompt $\mathbf{z^a}$, tissue affordance map $\mathbf{z}^\gamma$, and tempo-

ral tool-tip trajectory $\mathbf{z^p}$, enabling the model to generate temporally coherent surgical action videos with realistic tool-tissue interactions (Fig. 1a).

2.2 Lightweight Conditioning Signals: Language Prompting ($\mathbf{z^a}$): To incorporate tool-action context, we condition on language prompts via LTX's text mechanism, following the template: "A robotic da Vinci {tool} performs {action} during a cholecystectomy. The laparoscopic camera view shows the {tool} within the abdominal cavity. The {tool} moves precisely as it executes the {action} motion." Affordance Conditioning ($\mathbf{z}^\gamma$): To guide where tool-tissue interaction occurs, we condition the model on a 2D binary affordance map I_a defined with respect to the reference frame I_f. The affordance embedding $\mathbf{z}^\gamma$ is extracted via a VAE encoder. Tool Trajectory ($\mathbf{z^p}$): To precisely control instrument tip motion and tool-tissue interaction, we encode the tool trajectory as a temporal sequence of 2D maps $I_p^i \in \mathbb{R}^{h\times w\times 3}$, $i \in [0, k]$, where k is the sequence length. A trajectory point at time i is represented as a fixed-radius circle centered at the tool-tip (x, y) in I_p^i, with instrument class encoded through the R and G channels. Encoding the tool tip and its class directly in the RGB channels lets us reformulate our task as video-to-video diffusion without requiring major architectural changes. The embedding $\mathbf{z^p}$ is obtained by passing the sequence through a VAE encoder.

2.3 Depth Consistency Loss: In the presented formulation, all spatial conditioning signals are expressed as 2D inputs, with control in the Z-dimension learned implicitly by LTX during training. However, due to the presence of vital anatomy in the surgical scene, omitting depth guidance may result in inaccurate movements that violate critical safety constraints. This motivates the introduction of depth control during training, leading us to develop a novel loss for depth consistency ($\mathcal{L}_{\text{DC}}$). We first generate corresponding depth-mask videos for all RGB videos in our training dataset using Depth Anything V2 [29]. After encoding these depth videos into the latent space (Fig. 1b), we introduced a cross-attention layer and a projection head that learns to reconstruct masked depth latents. This reconstruction uses the denoised RGB latent tokens created after a forward pass of the diffusion transformer. The depth consistency loss (Fig. 1c), $\mathcal{L}_{\text{DC}}$, is then computed using a Smooth ℓ_1 loss, which follows an ℓ_2 loss for small errors and an ℓ_1 loss for large ones, providing robustness to outlier depth token predictions. Through this loss, we enforce geometric consistency in the Z-dimension without requiring explicit depth inputs at inference time. Unlike GeoVideo [1], which regularizes geometry using reprojected 3D depths, SAW uses a local Smooth-L1 loss on depth reconstructed from RGB video tokens, requiring no depth at inference.

3 Experiment and Results

3.1 Experimental Setup

Dataset: We custom-curate a dataset of 12,044 video clips from 101 surgical videos sourced from 21 Youtube videos (2,760 video clips) and 80 videos (12,878

video clips) taken from four publicly available datasets: 10 videos from HeiChole (1,794 video clips) [26], 30 videos from Cholec80 (6,042 video clips) [24], 32 videos from SurgVU (4,278 video clips) [33], and 8 videos from CRCD (764 video clips) [17]. Each video clip was segmented to capture a *single surgical action* and was manually annotated for *video-level action* (clipping, grasping, cutting, or dissecting) and *active tool usage* (grasper, hook, clipper, or scissors) as well as *tissue affordance*, which indicates the tissue region undergoing tool-tissue interaction. Frame-wise annotations are provided for *tool-tip positions* p = (x, y) in pixel coordinates. All video clips are preprocessed to remove black borders, letterboxing, and text overlay. We standardized all video clips to 81 frames at 25 fps, truncating longer videos and padding shorter ones by repeating the final frame. All videos were then processed at a resolution of 1024x576. The dataset is partitioned based on the 101 source videos, with the train set containing 11,502 video clips (clipping = 443, grasping = 1,964, cutting = 273, dissecting = 8,822), and the test set comprising 542 video clips (clipping = 29, grasping = 126, cutting = 20, dissecting = 367).

Training and Inference All our experiments are conducted using a single NVIDIA A100. We fine-tuned LTX using IC-LoRA for 7,500 steps, with $\alpha = 128$, $\text{lr} = 2 \times 10^{-4}$, AdamW optimizer, and bfloat16 mixed precision training. To improve alignment with groundtruth videos and enhance generation diversity, we employ classifier-free guidance with first-frame conditioning in 20% of cases. During inference, we use 50 denoising steps with a guidance scale = 3.5. All generated videos are 81 frames.

Evaluation Metrics We evaluate the visual quality of generated video using FVD [25], CD-FVD [9], SSIM [28], PSNR, and LPIPS [32] metrics. FVD assesses visual realism at the video level. Since frame-level perceptual metrics cannot directly assess the accuracy of tool-tissue interactions, we rely on CD-FVD, which is sensitive to temporal inconsistencies [9], to implicitly capture irregular tool-tissue motion. We further report SSIM, PSNR, and LPIPS to quantify frame-wise structural and perceptual image quality.

3.2 Results and Ablation Studies

Baseline comparison: Table 1 highlights the performance of our model compared to existing generative models. SAW achieves the lowest FVD at 224.28. Moreover, it excels at generating temporally consistent videos, demonstrated

Table 1: Quantitative evaluation of our model performance in video generation against state-of-the-art models on a held-out test set. WAN [27] and LTX_b [11] are used off-the-shelf whereas SurgSora [6] is fine-tuned on our training dataset.

Model	FVD ↓	CD-FVD ↓	SSIM ↑	PSNR ↑	LPIPS ↓
WAN [27]	439.60	429.67	0.5750	15.82	0.4480
LTX_b [11]	319.37	504.31	0.5630	15.07	0.4920
SurgSora [6]	541.61	546.82	0.4163	17.10	**0.3825**
Ours (SAW)	**224.28**	**199.19**	**0.5948**	**17.36**	0.4100

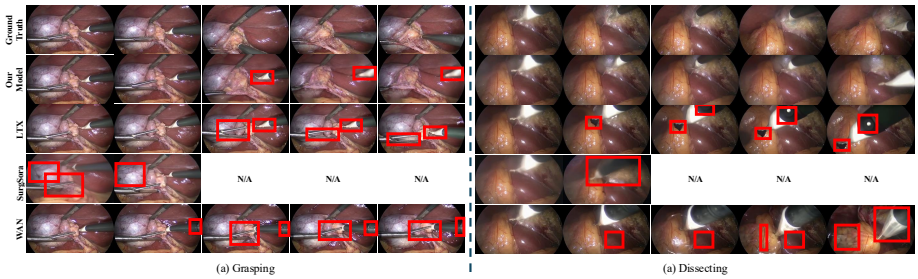

Fig. 2: Qualitative examples of generated surgical videos for **(a)** grasping and **(b)** dissecting actions. Red overlays indicate where a generated video lacks realism or becomes visually distorted in comparison to the ground truth video at the corresponding frame.

Table 2: Ablation study across conditioning modalities and loss components. Each row removes a single conditioning signal or loss term from the full model. Best results per metric are in **bold**.

Ablated Component	FVD ↓	CD-FVD ↓	LPIPS ↓	SSIM ↑	PSNR ↑
SAW	224.28	**199.19**	0.4100	0.5948	17.36
Ablation on Conditioning signals					
w/o Affordance	223.78	200.10	0.4101	0.5945	17.33
w/o First Frame	1096.21	338.75	0.6217	0.4848	12.50
w/o Language	219.23	204.89	0.4052	**0.5977**	**17.47**
w/o Trajectory	278.26	344.79	0.4325	0.5818	16.45
Model trained without depth consistency loss					
w/o $\mathcal{L}_{\text{DC}}$	**188.34**	207.59	**0.4017**	0.5922	17.37

through a CD-FVD of 199.19, outperforming the 546.82 of the also trajectory-conditioned surgical generative model SurgSora. For non-trajectory-conditioned models, the closest CD-FVD belongs to WAN with 429.67, suggesting the benefit of incorporating a large pretraining dataset for video generation. On frame-level content and structural metrics, SAW also outperforms all baselines with an SSIM of 0.5948 and PSNR of 17.36. SAW also demonstrated competitive frame-wise perceptual similarity with a LPIPS of 0.41. Fig. 2 illustrates the quality of the generated videos across two separate actions. Even in the presence of endoscopic artifacts, our model can produce a coherent video that reflects the underlying action, demonstrating its controllability, robustness, and realism.

Ablation Studies: We assess the impact of our design choices and conditioning strategies in Table 2. Removing trajectory conditioning significantly impacts performance, as indicated by a consistent decrease in all the evaluation metrics. Removing first-frame conditioning reduces visual quality, as indicated by drop performance in FVD, LPIPS, and SSIM. Furthermore, temporal consistency is affected as reflected by an increased CD-FVD to 338.75. Disabling the depth

consistency loss, $\mathcal{L}_{\text{DC}}$, primarily increases CD-FVD, suggesting a reduction in temporal consistency of tool and tissue movements. A paired bootstrap test on the test set confirms this improvement is significant (CD-FVD: 199.19 vs. 207.59, $p < 0.05$). Overall, SAW with all design choices is the most balanced model across all evaluation metrics.

4 Downstream Applications

We also demonstrate two downstream applications of our SAW model: (a) synthesizing rare surgical action videos to address class imbalance and enhance model performance on a real test set, and (b) exploring its potential to serve as a simulation engine for generating realistic tool–tissue interactions from surgical scenes and simulator-derived instrument trajectories.

4.1 Synthetic Video Generation for Action Recognition: With action recognition being a major research topic in surgical AI [20, 14], we design and conduct an experiment to demonstrate our framework's potential in augmenting rare action videos to improve model performance. Our train set exhibits a long-tail distribution, with cutting (273 clips) and clipping (443 clips) actions significantly underrepresented compared to dissecting (8,822 clips) and grasping (1,964 clips). Generating synthetic videos of rare events: Leveraging our controllable video generation framework, we generated synthetic videos of rare actions (cutting and clipping) to augment the training set. As Fig. 3b describes, we overlay segmented surgical tools using SAM3 [5] onto diverse background surgical scenes, and leverage the tissue affordance and tool trajectory extracted from the source background video –substituted with the target tool class – to condition generation of the augmented video. Superimposed surgical tools, which are acquired from existing training video clips, are blended into the background scene using Poisson image blending [18] to increase realism of the starting scene used to guide the diffusion process. While background scenes were diverse, we ensured relevance to each action – for example, clipping actions were paired with scenes containing blood vessels, increasing the diversity of tool–anatomy pairings. In total, we generated 287 synthetic videos: 110 clipping and 177 cutting. Experimental setup: We train two AI models – a spatiotemporal CNN [23] and a vision transformer (ViT) model [7] – for video action recognition. Consistent

Table 3: Per-class action recognition F1 score before and after video augmentation. Best results per action label are in **bold**.

Actions	Spatiotemporal CNN [23]		ViT [7]	
	w/o Aug	with Aug	w/o Aug	with Aug
Clipping	20.93	**43.14**	**84.62**	83.64
Cutting	0.00	**8.33**	30.77	**47.06**
Grasping	**58.52**	55.87	**74.19**	73.80
Dissection	83.87	**84.25**	86.27	**88.26**

with state-of-the-art approaches in surgical action prediction [20], both models take as input 16-frame segments from each video to predict the action. To evaluate our framework's utility in enhancing model performance through data augmentation, we train each model under two settings: (i) using video clips only from the original training set, and (ii) augmenting the training set with synthetically generated video clips of rare actions. Results and Discussion: Table 3 summarizes the performance of both models trained with and without synthetic rare action videos and evaluated on a held-out real test set. For both models, we observe that augmentation with synthetic videos can improve recognition of underrepresented actions (clipping and cutting) on real test videos, validating our framework's utility in generating realistic and effective training data. Since augmentation targets only the underrepresented classes, improvements are visible in these actions, while grasping and dissection remain largely unchanged.

4.2 Realistic Tool Tissue Interaction Engine for Surgical Simulation: While physics-based surgical simulators have made significant progress, they struggle to accurately model complex tissue-tool interactions and tissue deformation in real time [31]. Towards addressing this challenge, we explore our controllable surgical action video generation framework as a potential engine to model realistic tool-tissue interaction and tissue deformation in near real-time. Qualitative Experimental Setup: As shown in Fig. 3, we build a surgical simulator using Isaac lab [16], encompassing surgical tools in an empty background. Given a set of simulator-derived data (instrument segmentation, their corresponding tool tip trajectories, and tissue affordance), we demonstrate our framework's flexibility and controllability in generating surgical action videos – modeling tool-tissue interaction with realistic tissue deformations – with any given background surgical scene. However, this represents an initial proof-of-concept. Future work will address enforcing instrument joint kinematics, supporting more instruments and scenes, expert ratings of generated videos, and real-time inference.

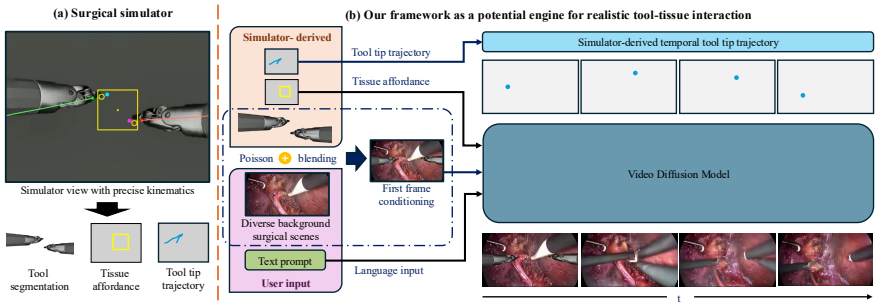

Fig. 3: **a)** Simulator-derived tool segmentations, tissue affordance, tool tip trajectory. **b)** Simulator-derived tools are overlayed on a background surgical scene and used as conditional inputs to our video diffusion model, allowing us to model both tool kinematics and background tissue deformation.

5 Conclusion

We present Surgical Action World (SAW) – a step toward surgical action world modeling through controllable and scalable video diffusion. By conditioning on four lightweight signals (language prompts, reference frame, tissue affordance mask, and tool-tip trajectories), SAW achieves state-of-the-art temporal consistency and visual quality while enabling scalable inference without expensive annotations. A depth consistency loss further encourages geometric coherence in tool motion, promoting plausible tool-tissue interactions that extend beyond purely 2D trajectory conditioning. We demonstrate the downstream utility of SAW for surgical AI through rare action augmentation that improves action recognition on real data, and for simulation through rendering tool-tissue interactions from simulator-derived trajectories. Future work will focus on improving the integration of affordance and language cues for richer scene understanding, extending to longer video generation, incorporating additional instruments and anatomical scenes, and achieving real-time inference for closed-loop simulation.

Disclosure of Interests. The authors have no competing interests.

Acknowledgments. This work was in part supported by the Johns Hopkins University SURPASS Grant, Malone Center Seed Grant, and Johns Hopkins University internal funds.

References

1. Bai, Y., Fang, S., Yu, C., Wang, F., Huang, Q.: Geovideo: Introducing geometric regularization into video generation model. Advances in Neural Information Processing Systems **38**, 57602–57622 (2026)
2. Biagini, D., Navab, N., Farshad, A.: Hierasurg: Hierarchy-aware diffusion model for surgical video generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 310–319. Springer (2025)
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
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: Forty-first International Conference on Machine Learning (2024)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chen, T., Yang, S., Wang, J., Bai, L., Ren, H., Zhou, L.: Surgsora: Object-aware diffusion model for controllable surgical video generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 521–531. Springer (2025)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

8. Fazzari, E., Stefanini, C.: Deep reinforcement learning for surgical robotics with state and image information: A survey (2026)
9. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in fréchet video distance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7277–7288 (2024)
10. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3), 440 (2018)
11. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
12. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)
13. 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)
14. Liao, W., Zhu, Y., Zhang, H., Wang, D., Zhang, L., Chen, T., Zhou, R., Ye, Z.: Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: a systematic review. Frontiers in Surgery **12**, 1551838 (2025)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
16. Mittal, M., Roth, P., Tigue, J., Richard, A., Zhang, O., Du, P., Serrano-Munoz, A., Yao, X., Zurbrügg, R., Rudin, N., et al.: Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. arXiv preprint arXiv:2511.04831 (2025)
17. Oh, K.H., Borgioli, L., Mangano, A., Valle, V., Di Pangrazio, M., Toti, F., Pozza, G., Ambrosini, L., Ducas, A., Žefran, M., et al.: Expanded comprehensive robotic cholecystectomy dataset (crcd). arXiv preprint arXiv:2412.12238 (2024)
18. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 577–582 (2023)
19. Scheikl, P.M., Tagliabue, E., Gyenes, B., Wagner, M., Dall'Alba, D., Fiorini, P., Mathis-Ullrich, F.: Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery. IEEE Robotics and Automation Letters **8**(2), 560–567 (2022)
20. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. International journal of computer assisted radiology and surgery **18**(6), 1053–1059 (2023)
21. Sivakumar, S.K., Frisch, Y., Ghazaei, G., Mukhopadhyay, A.: Sg2vid: Scene graphs enable fine-grained control for video synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 511–521. Springer (2025)
22. Tagliabue, E., Pore, A., Dall'Alba, D., Magnabosco, E., Piccinelli, M., Fiorini, P.: Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3261–3266. IEEE (2020)
23. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
24. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)

25. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
26. Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical image analysis **86**, 102770 (2023)
27. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 **3**(4), 6 (2025)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
29. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
30. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114 **1**(2), 6 (2023)
31. Zhang, J., Zhong, Y., Gu, C.: Deformable models for surgical simulation: a survey. IEEE reviews in biomedical engineering **11**, 143–164 (2017)
32. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
33. Zia, A., Berniker, M., Nespolo, R., Perreault, C., Wang, Z., Mueller, B., Schmidt, R., Bhattacharyya, K., Liu, X., Jarc, A.: Surgical visual understanding (surgvu) dataset. arXiv preprint arXiv:2501.09209 (2025)