

Exploring World Models for Surgical Video Synthesis for Tool Localisation and Tracking

Moucheng Xu^{1*}, Razvan Caramalau^{1*}, Justin Choi^{1,2}, Matthew Lee^{1,2},
Maria Robu¹, Imanol Luengo^{1,2}, and Danail Stoyanov^{1,2}

¹Medtronic Digital Technologies
London, UK

²UCL Hawkes Institute, University College London,
London, UK

* These authors contributed equally and share first authorship.
{moucheng.xu2, razvan.caramalau}@medtronic.com

Abstract. Data scarcity is a persistent bottleneck for surgical AI, particularly when new instruments are introduced and few annotated recordings are available. In this work we investigate whether modern video-generation “world models” can act as a synthetic data engine for surgical tool localisation and tracking. We study two open-source backbones from NVIDIA’s Cosmos family: the general-purpose Cosmos-Predict 2.5 and the domain-adapted Cosmos-H-Surgical, and evaluate each on a curated subset of CholecTrack20 under three regimes: zero-shot inference, LoRA fine-tuning and full supervised fine-tuning. Generation is performed in Video2World mode, conditioned on the first two frames of each clip and a structured text prompt describing surgery type, phase, tools and tool tracks. We evaluate the resulting clips along two complementary axes: video quality (FVD, LPIPS and three VBench dimensions) and domain quality for downstream perception, measured by running a YOLO-v7 tool detector on both real and generated videos and comparing bounding-box consistency and per-frame track adherence at 10 fps. Across 367 test clips, Cosmos-H-Surgical achieves the better distributional realism while Cosmos-Predict 2.5 offers the better paired perceptual fidelity and the strongest tool-track reproduction. Overall, world models can plausibly render new surgical scenes, but faithfully reproducing instrument geometry frame-by-frame remains substantially harder than matching global video statistics.

Keywords: Surgical video synthesis · World models · Cosmos-Predict · Surgical tool localisation · Synthetic data · CholecTrack20.

1 Introduction

Artificial intelligence (AI) systems require large volumes of high-quality data to achieve robust performance, yet acquiring and curating such data in surgical domains remains challenging. When a new instrument is introduced intraoperatively, collecting representative samples typically demands additional surgeon

involvement or extensive post hoc analysis of recordings. This early-deployment scarcity limits model performance and reduces generalization to novel scenes and tools. Recent advances in Generative AI have improved the realism, physical consistency, and temporal coherence of synthesized scenes, increasingly through large-scale “world models” [2, 3, 9] trained on massive video datasets. Motivated by this paradigm, we adopt Cosmos-Predict 2.5 [1], a 2B-parameter flow-based model from NVIDIA’s Cosmos family that supports Text2World, Image2World, and Video2World generation. As a domain-adapted baseline, we also use Cosmos-H-Surgical [4], post-trained from Cosmos 2.5 with LoRA [5] on over 300k surgical frames, which is expected to have a smaller domain gap to our target dataset.

We study surgical video synthesis on a variant of CholecTrack20 [8]. Both world models generate clips in Video2World mode, conditioned on the first two frames of the original clip, under three setups: zero-shot inference, LoRA training, and supervised fine-tuning (SFT). Our study has two parts: a qualitative appreciation of the generated videos, and an empirical evaluation of whether text-prompt control can place tools on trajectories similar to the original. The goal is to enable re-generating data from similar surgical scenes but with different tool trajectories. Our findings show that Cosmos-H-Surgical SFT yields the best scene realism, whereas Cosmos 2.5 SFT captures the best text conditioning for tool bounding-box matching. This suggests that models already post-trained on surgical data bridge the domain gap but begin to lose the physics and control capabilities of the base model. Our work opens possibilities for better action control and for estimating the data required for reliable surgical video synthesis.

2 Related Work

COSMOS / COSMOS-H-Surgical NVIDIA’s COSMOS-predict2.5 is an open-source physical AI world model trained on over 200 million clips [1, 2]. Like Genie3, it generates future frames from multi-modal controls, but focuses on physical worlds, especially for robotics and autonomous driving. Its empirical success and open-source nature motivate our adoption of it. COSMOS-H-Surgical [4] extends COSMOS2.5 using over 300k curated frames of surgical robotic tasks (e.g. needle grasping, puncture, suture pulling, knotting). As the variant closest to our target domain, we adopt it as well.

Surg- Σ Surg- Σ [12] is a recent large multi-modal dataset (1.58M images, 4.21M conversations) for surgical AI foundation models. However, its dense-prediction labels lack instance-level tool bounding boxes, a common gap in the field that limits downstream use in surgical tool tracking. This motivates our exploration of generative foundation models to synthesize tool bounding boxes for training specialized tracking agents.

3 Methods

Our synthetic data engine pipeline (Figure 1) leverages the existing Cosmos-predict2.5-style world models for generating synthetic data for surgical tool localisation and tracking. We use tailored text prompts to guide the synthesis of

the surgical video containing the tools. For each video, the prompt includes the following information: surgery types, clip temporal information, surgery phase, video conditions, tool types, tool tracks in bounding boxes format.

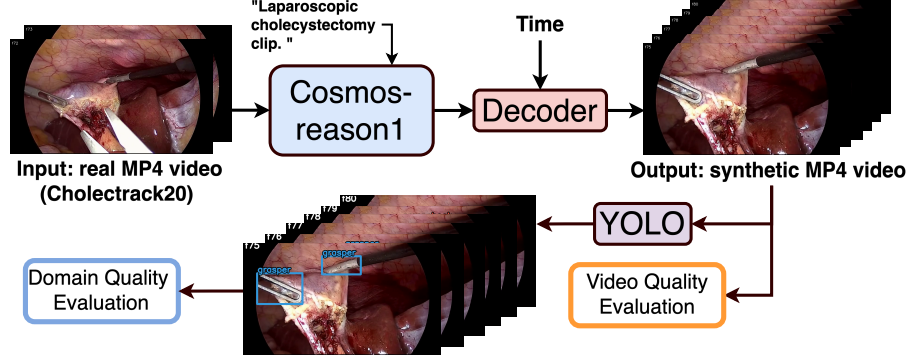


Fig. 1: The proposed data engine pipeline. Cosmos-reason1 and the DiT based VLM Decoder are the core components of Cosmos-predict2.5 and Cosmos-H-Surgical. A YOLO v7 is used to assist the evaluation of the domain quality of the synthetic videos.

Data CholecTrack20 [8] contains 20 videos for laparoscopic cholecystectomy for multi-class multi-tool tracking. We noticed that in the provided data of CholecTrack20, there are no source mp4 videos available in their official training split and their validation split, whereas COSMOS-predict2.5 supports mp4 video inputs. Therefore, we are limited to use their testing dataset which contains 8 mp4 videos. However, one of the testing videos was corrupted during the downloading. Eventually, we ended up with total 7 useable mp4 videos. Among the 7 available mp4 videos, we randomly selected 4 videos for training, 1 video for validation, and the rest 2 videos for testing. Then we further pre-processed the mp4 data into the format suitable for the training and inference of COSMOS-predict2.5. Our pre-processing includes, chunking each video to non-overlapping 10 seconds clips, each clip is 10FPS, the resolution has been kept as the raw videos at 480x854. In the end, we have 927 training clips, 86 validation clips and 367 testing clips. We as well converted their provided surgery descriptions in their json format into txt based prompts.

Post-training We post-trained two versions of COSMOS, namely, COSMOS-predict2.5 and COSMOS-H-Surgical. For each variant of the COSMOS, we explored both supervised fine-tuning (SFT) and low-rank adaption fine-tuning (LoRA) post-training approaches. We used 4 NVIDIA A100 GPUs for the training and the inference of COSMOS-H-Surgical, and we used 4 NVIDIA RTX A6000 GPUs for the training and the inference of the COSMOS-predict2.5.

We post-trained each model with batch size 1 for 10 epochs and we saved the checkpoints after every epoch. We did not train the models beyond 10 epochs,

Table 1: Video-quality evaluation of the generated surgical videos on the Cholec-Track8 split-2 test set (367 clips). FVD and LPIPS are computed against the paired ground-truth clips; the remaining three are VBench [6] computed on the generated clips alone. Arrows denote whether higher ($\uparrow$) or lower ($\downarrow$) is better; best value per column in **bold**.

Model	FVD $\downarrow$	LPIPS $\downarrow$	Dyn. Degree $\uparrow$	Img. Quality $\uparrow$	Overall Consist. $\uparrow$
Cosmos 2.5 (0-shot)	5821	0.833	0.054	0.513	0.153
Cosmos 2.5 (LoRA)	568	0.466	0.828	0.688	0.126
Cosmos 2.5 (SFT)	340	0.425	0.788	0.691	0.120
Cosmos-H-S(0-shot)	2763	0.574	0.411	0.590	0.137
Cosmos-H-S(LoRA)	2779	0.574	0.352	0.590	0.137
Cosmos-H-S (SFT)	333	0.459	1.000	0.676	0.120

because we observed that the validation samples started to have unrealistic hallucinations across all of the trainings. The other training hyper-parameters are kept as the default values as the provided COSMOS-predict2.5 from NVIDIA.

Surgical Tool Localisation In evaluation, in order to assess the domain quality of the generated surgical videos for the downstream tool localisation tasks, we need to infer the bounding boxes from the generated videos. To achieve this, we adapted a popular open-sourced toolkit Ultralytics [10] as our AI agent to localise the surgical tools from the generated videos. More precisely, we used a pre-trained medium variant of Yolo-v7, because of its good trade-off between performances and inference costs. The extracted bounding boxes of the tools from the synthetic videos are then further analysed to calculate the metrics to measure the domain quality.

4 Results

4.1 Video Quality Score

We evaluate the generated video quality with the following metrics: Fréchet Video Distance (FVD) [11], Learned Perceptual Image Patch Similarity (LPIPS) [13], VBench dimensions (dynamic degree, image quality, overall consistency) [7]. Table 1 shows that fine-tuning is consistently beneficial. Cosmos 2.5 zero-shot is far off the target data manifold (FVD 5821, LPIPS 0.833) and almost static (dynamic degree 0.054). Although zero-shot attains the highest overall consistency (0.153), it is worthy to point out that this is an artifact of its near-frozen output rather than genuine prompt adherence and should be read alongside its poor fidelity. Cosmos-H Surgical, which is pre-trained on surgical data, has much better zero-shot performance (FVD 2763, LPIPS 0.574). Adaptation sharply improves the Cosmos 2.5 backbone: LoRA cuts FVD to 568 and LPIPS to 0.466 while injecting realistic motion (dynamic degree 0.828), and SFT yields the best LPIPS (0.425) and imaging quality (0.691) with a very low FVD (340). For the Cosmos-H Surgical backbone, LoRA with the default setting barely changes any

metric relative to zero-shot, whereas SFT transforms it into the best model on distributional realism, reaching the lowest FVD overall (333) and competitive LPIPS (0.459) and imaging quality (0.676). Its dynamic-degree score of exactly 1.000 means *every* clip is flagged as dynamic; while this confirms the model no longer produces static videos, a saturated value warrants caution, as it may also reflect excessive or unnatural motion that VBench’s binary motion test cannot distinguish from plausible surgical movement. Overall consistency is low and tightly clustered (0.12-0.15) across all adapted models and is not discriminative here. Taken together, the two SFT models are the strongest: Cosmos 2.5 (SFT) offers the best paired perceptual fidelity (LPIPS) and imaging quality, while Cosmos-H Surgical (SFT) offers the best distributional realism (FVD) and the most motion, making them the two configurations we carry forward to the detection- and track-consistency analyses. The trends in the quantitative results are also reflected in the qualitative results as shown in Figure 2.

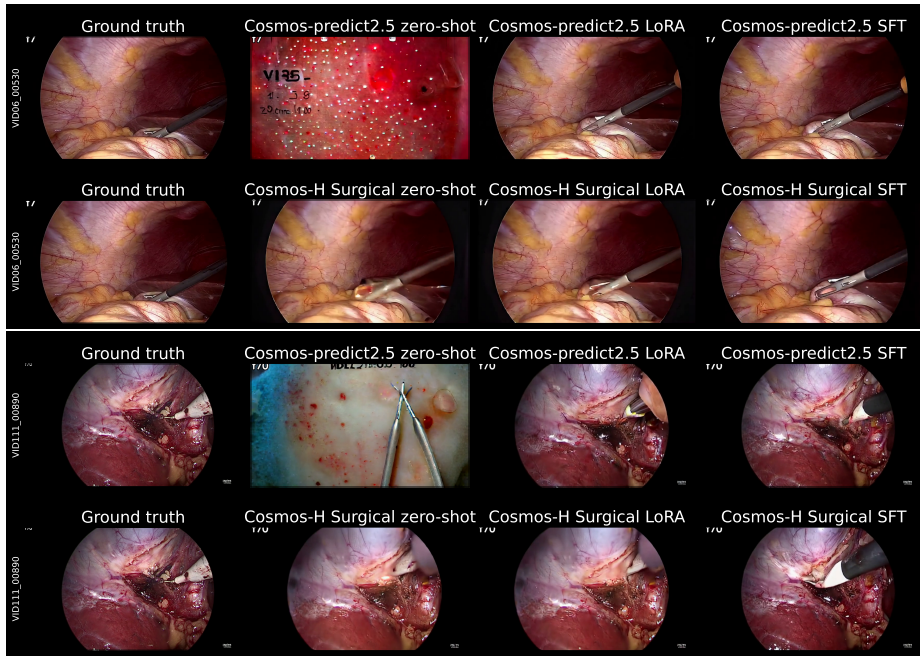


Fig. 2: Visual comparisons between ground truth frames and the generated frames. The frames were generated conditioning on the first 2 ground truth frames. *Row 1-2*: 5-th frame from a clip in VID06. *Row 3-4*: 70-th frame from a clip in VID111.

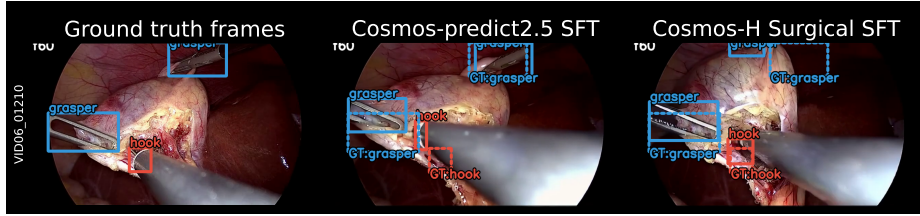


Fig. 3: Detected instruments by YOLOv7 in the ground truth frames and the generated frames. Solid bounding boxes are the detected bounding boxes on the frames. Dotted bounding boxes are the bounding boxes detected in the ground truth frames, for easier visual comparisons.

Table 2: Object-detection consistency of generated surgical videos on the Cholec-Track8 test set. Detections are produced by the same YOLO model on each set of generated clips and matched ($\text{IoU} \geq 0.5$) against detections on the corresponding ground truth videos, which serve as the reference. All metrics are aggregated over 367 videos. Arrows denote whether higher ($\uparrow$) is better; best value per column in **bold**.

Model	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	mIoU $\uparrow$	mAP@0.5 $\uparrow$
Cosmos 2.5 (zero-shot)	0.175	0.033	0.056	0.893	0.021
Cosmos 2.5 (LoRA)	0.081	0.083	0.082	0.758	0.028
Cosmos 2.5 (SFT)	0.106	0.103	0.105	0.759	0.035
Cosmos-H Surgical (zero-shot)	0.144	0.066	0.091	0.784	0.031
Cosmos-H Surgical (LoRA)	0.143	0.067	0.091	0.781	0.032
Cosmos-H Surgical (SFT)	0.055	0.058	0.056	0.698	0.014

4.2 Domain Quality Score: Tool Localisations

We treat the tool detections on the original videos as ground truth. We only consider the synthetic frames that contain the same instruments, in the same places, as the real surgery. In order to do so, for every video we match predicted and reference boxes frame-by-frame within each class using a greedy, highest-IoU assignment at an IoU threshold of 0.5, from which we compute precision, recall, F1, the mean IoU of matched boxes, and a per-video average precision (mAP@0.5) averaged across classes and videos (Table 2). Overall, we confirm that it remains a challenging task to produce synthetic frames that contain frame-by-frame comparable surgical tools. However, we as well confirm that supervised adaptation is still consistently more beneficial than zero-shot at improving the synthetic tool fidelity, as reflected in the detection metrics that, SFT achieves the best recall (0.103), F1 (0.105), and mAP (0.035), roughly tripling the recall of the zero-shot model. Especially, the SFT Cosmos 2.5 model offers the best trade-off between detection recall and precision.

Table 3: Track adherence of generated surgical videos, measured against the original-video YOLO detections used as a dense 10-fps reference on the CholecTrack8 split-2 test set (367 videos). Every reference detection is matched to the best same-class detection at the same frame in the generated video. Arrows indicate whether higher ($\uparrow$) or lower ($\downarrow$) is better; best value per column in **bold**.

Model	Track IoU $\uparrow$	Presence rate $\uparrow$	Center error $\downarrow$
Cosmos 2.5 (zero-shot)	0.054	0.081	0.056
Cosmos 2.5 (LoRA)	0.178	0.493	0.101
Cosmos 2.5 (SFT)	0.207	0.558	0.098
Cosmos-H Surgical (zero-shot)	0.119	0.279	0.088
Cosmos-H Surgical (LoRA)	0.119	0.280	0.089
Cosmos-H Surgical (SFT)	0.150	0.543	0.123

4.3 Domain Quality Score: Tool Trajectories

We treat the YOLO detections on the ground-truth clips as reference tracks and the detections on each model’s generated clip as predictions, both sampled at the native 10 fps (100 frames per clip), with each detection expressed in normalized coordinates $[t_{lx}, t_{ly}, w, h]$. For every reference detection at frame f with class c , we take the highest-IoU same-class prediction at the same frame and report three quantities: (i) **Track IoU**, the mean best-match IoU (unmatched references contribute 0, penalizing both misplacement and missing instruments); (ii) **Presence rate**, the fraction of reference detections with any same-class overlap (IoU > 0); and (iii) **Center error**, the mean box-center distance over positively matched detections only. Because the reference is dense and drawn from the ground-truth videos, these metrics measure per-frame spatial fidelity of the generated tracks.

Table 3 shows that track fidelity benefits consistently from supervised adaptation. Cosmos 2.5 zero-shot is effectively ungrounded, recovering only 8.1% of reference detections at a Track IoU of 0.054; its low center error (0.056) is an artifact of emitting so few instruments and should not be read as superior localization. Adaptation yields large gains: LoRA lifts presence to 0.493 and IoU to 0.178, while SFT is strongest overall, matching 55.8% of detections at a Track IoU of 0.207 with a competitive center error of 0.098. For Cosmos-H Surgical, zero-shot and LoRA are nearly identical (IoU $\approx$ 0.119, presence $\approx$ 0.28), showing LoRA adds little on top of surgical pre-training, whereas SFT roughly doubles presence to 0.543 but worsens center error to 0.123, the highest of all models. This “high-presence, low-precision” behaviour mirrors the false-positive inflation seen in the detection-consistency analysis: it reliably paints instruments in but not along the true trajectories. Overall, the fully SFT Cosmos 2.5 model best reproduces the original tracks.

Figure 4 contrasts a well-matched (left) and a random poorly-matched (right) clip for SFT Cosmos 2.5, both chosen among the highest-motion clips. In the good example (presence 0.72), the generated grasper and hook tracks (dashed)

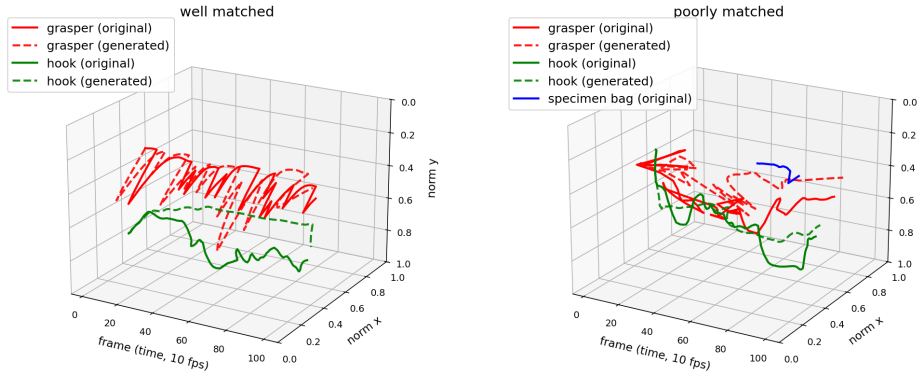


Fig. 4: Qualitative track adherence for the SFT Cosmos 2.5 model (COSMOS 2.5 SFT). Each panel shows 3D instrument-center tracks with time (frame index at 10 fps) and the normalized image x/y ; solid lines are the original-video detections (reference) and dashed lines the detections on the generated video, and every curve is exponential-moving-average(EMA) smoothed to expose the underlying motion. *Left*: a well-matched, high-motion clip (Track IoU 0.37, presence 0.72). *Right*: a random poorly-matched, high-motion clip (Track IoU 0.10, presence 0.43).

follow the reference (solid) in x and y even under rapid motion, consistent with a Track IoU of 0.37. The poor example exposes the dominant failure mode: the generated tracks diverge (Track IoU 0.10) and the specimen bag (blue) is not reproduced at all. This shows even the strongest model can place instruments inconsistently despite plausible per-frame appearance, explaining why aggregate Track IoU remains modest. Visual realism and precise track adherence are related but distinct, and the latter is harder to satisfy.

5 Conclusion

We investigated whether open-source video world models (Cosmos-predict2.5 derived models) can serve as a synthetic data engine for surgical tool localisation, under different training configurations. Our results show that supervised fine-tuning yields the best distributional realism (FVD 333 for Cosmos-H-Surgical) and the strongest tool-track reproduction (Track IoU 0.207, presence 0.558 for Cosmos-Predict 2.5), while LoRA adds little on top of a backbone that is already surgically pre-trained. However, even the best configurations exhibit low absolute detection scores and a “high-presence, low-precision” failure mode, indicating that video-level realism does not yet translate into faithful frame-by-frame instrument geometry. Future work will focus on scaling and stronger spatial conditioning (e.g. bounding-box or mask guidance) and evaluating whether the generated clips actually improve a downstream identity-aware tracker.

6 Acknowledgment

Competing Interests: MX, RC, MT, MR, IL and DS are employees of Medtronic Digital Technologies. JC contributed to this work as part of their MSc dissertation conducted as a Research Intern at Medtronic Digital Technologies.

References

1. &, N., Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., Chattopadhyay, P., Chen, M., Chen, Y., Chen, Y., Cheng, S., Cui, Y., Diamond, J., Ding, Y., Fan, J., Fan, L., Feng, L., Ferroni, F., Fidler, S., Fu, X., Gao, R., Ge, Y., Gu, J., Gupta, A., Gururani, S., Hanafi, I.E., Hassani, A., Hao, Z., Huffman, J., Jang, J., Jannaty, P., Kautz, J., Lam, G., Li, X., Li, Z., Liao, M., Lin, C.H., Lin, T.Y., Lin, Y.C., Ling, H., Liu, M.Y., Liu, X., Lu, Y., Luo, A., Ma, Q., Mao, H., Mo, K., Nah, S., Narang, Y., Panaskar, A., Pavao, L., Pham, T., Ramezanali, M., Reda, F., Reed, S., Ren, X., Shao, H., Shen, Y., Shi, S., Song, S., Stefaniak, B., Sun, S., Tang, S., Tasmeen, S., Tchapmi, L., Tseng, W.C., Varghese, J., Wang, A.Z., Wang, H., Wang, H., Wang, T.C., Wei, F., Xu, J., Yang, D., Yang, X., Ye, H., Ye, S., Zeng, X., Zhang, J., Zhang, Q., Zheng, K., Zhu, A., Zhu, Y.: World simulation with video foundation models for physical ai (2026), <https://arxiv.org/abs/2511.00062>
2. Agarwal, N.N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., Dworakowski, D., Fan, J., Fenzi, M., Ferroni, F., Fidler, S., Fox, D., Ge, S., Ge, Y., Gu, J., Gururani, S., He, E., Huang, J., Huffman, J., Jannaty, P., Jin, J., Kim, S.W., Klár, G., Lam, G., Lan, S., Leal-Taixe, L., Li, A., Li, Z., Lin, C.H., Lin, T.Y., Ling, H., Liu, M.Y., Liu, X., Luo, A., Ma, Q., Mao, H., Mo, K., Mousavian, A., Nah, S., Niverty, S., Page, D., Paschalidou, D., Patel, Z., Pavao, L., Ramezanali, M., Reda, F., Ren, X., Sabavat, V.R.N., Schmerling, E., Shi, S., Stefaniak, B., Tang, S., Tchapmi, L., Tredak, P., Tseng, W.C., Varghese, J., Wang, H., Wang, H., Wang, H., Wang, T.C., Wei, F., Wei, X., Wu, J.Z., Xu, J., Yang, W., Yen-Chen, L., Zeng, X., Zeng, Y., Zhang, J., Zhang, Q., Zhang, Y., Zhao, Q., Zolkowski, A.: Cosmos world foundation model platform for physical ai (2025), <https://arxiv.org/abs/2501.03575>
3. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., Aytar, Y., Bechtle, S.M.E., Behbahani, F., Chan, S.C., Heess, N., Gonzalez, L., Osindero, S., Ozair, S., Reed, S., Zhang, J., Zolna, K., Clune, J., Freitas, N.D., Singh, S., Rocktäschel, T.: Genie: Generative interactive environments. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) *Proceedings of the 41st International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 235, pp. 4603–4623. PMLR (21–27 Jul 2024), <https://proceedings.mlr.press/v235/bruce24a.html>
4. He, Y., Guo, P., Xu, M., Li, Z., Myronenko, A., Imans, D., Liu, B., Yang, D., Gu, M., Ji, Y., Jin, Y., Zhao, R., Shen, B., Xu, D.: Cosmos-h-surgical: Learning surgical robot policies from videos via world modeling (2026), <https://arxiv.org/abs/2512.23162>
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), <https://arxiv.org/abs/2106.09685>

6. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models (2023), <https://arxiv.org/abs/2311.17982>
7. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models (2023), <https://arxiv.org/abs/2311.17982>
8. Nwoye, C.I., Elgohary, K., Srinivas, A., Zaid, F., Lavanchy, J.L., Padoy, N.: Cholec-track20: A multi-perspective tracking dataset for surgical tools. In: CVPR (2026)
9. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving (2025), <https://arxiv.org/abs/2503.20523>
10. Ultralytics: Ultralytics: Yolo and computer vision models. <https://github.com/ultralytics/ultralytics> (2026), gitHub repository, accessed 2026-06-23
11. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges (2019), <https://arxiv.org/abs/1812.01717>
12. Zeng, Z., Xu, M., Jiang, J., Guo, P., Xu, Y., Zhuo, Z., Low, C.H., He, Y., Yang, D., Lin, C., Gu, Y., Guo, J., Ban, Y., Xu, D., Dou, Q., Jin, Y.: Surg σ : A spectrum of large-scale multimodal data and foundation models for surgical intelligence (2026), <https://arxiv.org/abs/2603.16822>
13. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018), <https://arxiv.org/abs/1801.03924>