SurgGenesis: A Generative Surgical World Model for Future-Aware Surgical Understanding

Qirui Zhong[1], Quande Liu[2], Yang Chen[1], and Cheng Xue[1]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China
cxue@seu.edu.cn
[2] The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China

Abstract. Surgical scene understanding is constrained by data scarcity and long-tail class imbalance, as annotating surgical videos demands scarce clinical expertise while rare procedural events remain underrepresented. We tackle this through data generation and propose SurgGenesis, a generative surgical world model that learns future surgical state transitions in a latent space from visual observations and textual procedural descriptions, serving as a scalable surgical data engine. To adapt a large-scale video diffusion model to the data-scarce surgical domain without catastrophic forgetting, we introduce a three-stage progressive LoRA strategy that incrementally injects endoscopic visual priors, procedural semantics, and spatial motion cues. Rather than judging generation by visual fidelity alone, we validate SurgGenesis along two axes that separate a world model from a generic generator: (i) future-awareness, by comparing predicted futures against the real future through a GT-latent prediction gap and temporal-horizon analysis, and (ii) action-faithful usefulness, by checking that generated clips contain the commanded surgical triplet and that using them to rebalance long-tail events and drive downstream understanding (segmentation, triplet recognition, 3D reconstruction) yields action-specific gains. Conditioned futures therefore complement conventional long-tail sampling.
Code: https://zh-qr.github.io/SurgGenesis.

Keywords: Surgical World Model · Text-Controllable Video Generation · State-Transition Dynamics · Multimodal Representation Learning

1 Introduction

Surgical video models can synthesize realistic clips, but visual realism alone is insufficient for a surgical world model. The model must start from an observed operative context, respond to a procedural description, and produce a coherent future. This is demanding in endoscopy, where tool–tissue interactions, viewpoint motion, and procedural phase evolve together.

Representative surgical-vision systems adapt multimodal representations or track instruments from observed videos[1–3]. Related multimodal surgical-video efforts target clinically interpretable anatomy and device–tissue interaction[4]. Related medical world-model work has modeled latent disease trajectories for

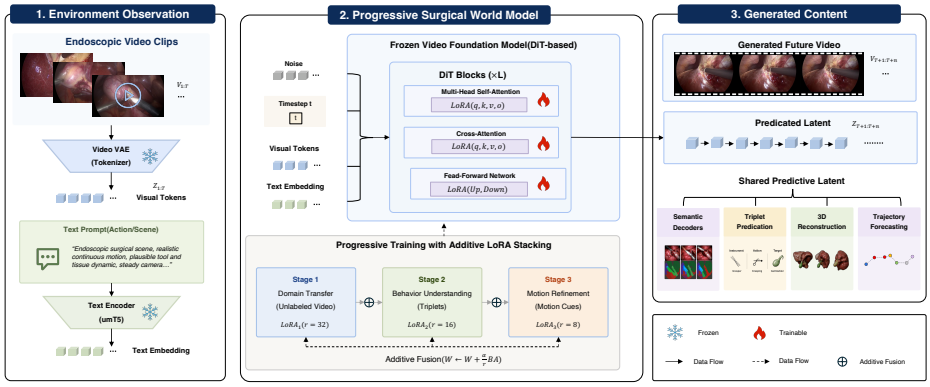

Fig. 1. Overview of SurgGenesis. (1) Visual observations and structured procedural text condition the model. (2) A three-stage strategy fine-tunes the foundational video world model. (3) Predicted latent states are decoded into future observations.

prediction and planning[5]; SurgGenesis brings this state-transition perspective to text-controlled surgical rollouts.

We use surgical world model operationally: given past observations and a structured procedure prompt, the model represents a conditional state transition and rolls out its consequences. A visually plausible sample may still change the operative state or ignore the requested action. The claim therefore requires that the prediction match the real continuation of its context and that the prompt controls the intended action. Generative world models provide a data-driven route to such spatio-temporal modeling[6, 7]; surgery is a demanding instance[8–10].

Generic video-customization methods improve controllability by disentangling subject or content representations from motion and by reducing interference among targeted LoRA adaptations[11, 12]. **SurgGenesis** models the conditional future transition $p(\hat{Z}_{t+1:T+n} \mid Z_{1:T}, c)$, where c is a structured textual description of procedure phase and instrument–verb–target (IVT) state. Unlike a generator assessed only by clip realism[13–15], SurgGenesis is evaluated as a context-conditioned, text-controllable rollout model. A **Three-Stage Progressive LoRA Fine-Tuning Strategy** adapts a large video diffusion model to endoscopic appearance, procedural semantics, and spatial dynamics without catastrophic forgetting. Additive LoRA stacking and rehearsal preserve earlier capabilities while later stages specialize the state-transition model.

Controlled rollouts can create rare, context-specific examples for costly, long-tailed surgical datasets[16–18]. Data scarcity and imbalance are application settings, not the sole claim: synthetic clips are useful only when they preserve the supplied context and requested action.

Our evaluation separates model properties from application utility. Fidelity metrics quantify visual realism. The Prediction Gap and multi-horizon RMSE test whether a rollout matches its real continuation. Thus, future awareness

concerns state transitions rather than data scarcity. A real-data recognizer tests whether the commanded IVT appears in the generated future. Segmentation, IVT recognition, 3D reconstruction, and long-tail augmentation then assess utility. Thus, a downstream gain or FVD improvement alone is not presented as evidence of a world model.

The main contributions of this paper are summarized as follows:

- We introduce **SurgGenesis**, a text-controllable generative surgical world model that rolls out future videos from observed surgical context and structured prompts containing procedure phase and IVT state.
- We propose a three-stage progressive LoRA adaptation strategy with additive stacking and data rehearsal, progressively incorporating endoscopic visual priors, procedural semantics, and spatial dynamics while mitigating catastrophic forgetting.
- We introduce a layered evaluation: fidelity metrics assess visual quality; the Prediction Gap and temporal horizons test context-to-future consistency; an IVT recognizer tests action faithfulness; and downstream tasks assess application utility.

2 Methodology

An overview of the proposed SurgGenesis framework is shown in Fig. 1. SurgGenesis learns predictive representations of surgical dynamics by integrating visual observations with structured procedural descriptions in a shared latent space. The resulting model rolls out a future surgical state that is conditioned on both the current scene and a user-specified procedure state.

2.1 Problem Formulation

We formulate surgical scene forecasting as the prediction of future latent states conditioned on historical observations and a structured prompt. Given a sequence of T past frames, denoted as $V_{1:T}$, the model predicts the next n latent states $\hat{Z}_{T+1:T+n}$ rather than generating pixels directly. Formally, it learns the conditional future distribution $p(\hat{Z}_{T+1:T+n} \mid Z_{1:T}, c)$, where c contains procedural text and coordinate-aware motion cues, and generates the future latent clip jointly through conditional flow matching. The text prompt is deliberately structured so that the phase, instrument, verb, and target can be changed independently: "Endoscopic surgical scene. Laparoscopic cholecystectomy, phase: {*phase*}. At frame 0, the {*instrument*} is {*verb*} the {*target*}. Predict the next 60 frames with smooth instrument motion, stable endoscopic viewpoint, and anatomically consistent tool–tissue interaction." For example, the fields can specify gallbladder extraction, a grasper, retracting, and the cystic plate. This representation turns the procedure description into an intervention-like control signal while retaining the observed scene as state context. The prompt controls the intended procedural transition, whereas the input frames anchor the anatomy, tool configuration,

illumination, and endoscopic viewpoint. Its role is therefore not unconstrained free-form synthesis: phase constrains which IVT combinations are meaningful, and the generated continuation must remain compatible with the visual state at frame 0. This separation is central to the proposed world-model setting. A generic text-to-video generator may satisfy the textual description while changing the scene arbitrarily; SurgGenesis must preserve the scene state while rolling out the requested action. We operationalize future awareness as context-to-future consistency: predicted latents and decoded frames should match the corresponding real continuation, not merely resemble surgery. Direct latent RMSE, decoded frame-space RMSE, and the RT-DETR trajectory readout are distinct diagnostics. The representation is therefore trained to capture spatio-temporal regularities such as instrument geometry, viewpoint continuity, and tool–tissue motion; segmentation and multi-view 3D reconstruction probe whether these rollouts preserve structural and geometric information.

2.2 Predictive Surgical World Modeling via Progressive Training

Surgical procedures evolve as complex and dynamic processes. However, most existing approaches focus on static interpretation of observed frames, which limits their ability to model future surgical states. To address this limitation, we develop a predictive surgical world model through **Progressive Surgical Dynamics Training**, a multi-stage paradigm designed to progressively incorporate domain-specific knowledge into a latent diffusion framework. The core of the model lies in the high-dimensional latent space Z introduced above, where a Diffusion Transformer (DiT) learns predictive representations of surgical dynamics by modeling the temporal evolution of surgical scenes. Following DiffSynth-Studio, training uses its native flow-matching SFT objective rather than direct L_1 latent regression: $\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{Z_0,\epsilon,t}[w(t)\|v_\theta(Z_t,t,c) - v_t^\star\|_2^2]$, where the scheduler constructs noisy latent Z_t and flow target $v_t^\star$ from clean future latent Z_0, Gaussian noise ϵ, and sampled timestep t. The first-frame latent is clamped as conditioning and excluded from the loss.

Stage 1: Domain Transfer. The model is initially adapted to the surgical domain using large-scale unlabeled video clips (Cholec80) and fixed textual descriptions. Low-Rank Adaptation (LoRA) is applied across all 30 transformer blocks to effectively capture surgical visual priors, such as specific tissue textures and illumination, without requiring dense manual annotations.

Stage 2: Behavior Understanding. Procedural knowledge is injected by aligning visual transitions with structured prompts containing phase and instrument–verb–target fields. These fields can be varied at inference time while the observed frames retain the operative context. To prevent the catastrophic forgetting of Stage 1 capabilities, we introduce *Additive LoRA Stacking*: the learned LoRA weights from Stage 1 are permanently fused into the base model weights via calculated increments. A new LoRA is then initialized and restricted exclusively to the last 20 transformer blocks, reducing plasticity to protect early-layer feature extraction. Furthermore, a 20% data rehearsal strategy is employed,

mixing baseline dataset samples into the current training batches to maintain foundational stability.

Stage 3: Spatial Dynamics Refinement. We introduce coordinate-aware descriptions of instrument motion to improve spatial precision and physical plausibility. Following the additive stacking principle, Stage 2 weights are fused, and the final LoRA targets only the deepest 10 blocks with a highly conservative learning rate. This ensures the model learns fine-grained spatial trajectories without disrupting the harmonized text distribution and procedural semantics acquired in previous stages.

2.3 Layered Evaluation and Downstream Verification

Our evaluation separates controlled world-model behavior from application utility. The Prediction Gap isolates forecasting error from decoder capacity by comparing predicted-future decoding against an upper bound from Ground-Truth (GT) latents. FVD-style and LPIPS metrics characterize visual or distributional fidelity; the Prediction Gap and temporal horizons characterize context-to-future consistency; and the IVT recognizer characterizes prompt-level action faithfulness. Downstream tasks then test whether the generated rollouts are useful, rather than treating any single downstream gain as proof of a world model.

The application-level protocol evaluates augmentation, segmentation, triplet recognition, and 3D reconstruction. These tasks probe utility, structure, semantics, and geometry.

3 Experiments

3.1 Datasets

Experiments use the Cholec-series datasets. Cholec80[19], CholecT50[20], and CholecTrack20[21] support Stages 1–3, respectively. Segmentation uses CholecSeg8k[18] and Endoscapes2023[17]; triplet recognition and augmentation use CholecT50; and 3D reconstruction uses EndoNeRF[22]. All downstream benchmarks follow official partitions and standard protocols. Generation and internal evaluation use held-out clips with active instrument use. In the official five-fold CholecT50 protocol, conditioning, synthesis, and downstream training use only each fold's training partition; validation and test remain unchanged.

3.2 Implementation Details

All experiments are conducted on NVIDIA A100 GPUs. We adopt the dense WAN2.2-TI2V-5B[23] as the foundational world model and fine-tune it with progressive LoRA in DiffSynth-Studio. Its unified text/image-to-video formulation and high-compression WAN2.2-VAE provide the backbone for surgical future generation. The progressive LoRA stacking uses ranks 32/16/8 over the last 30/20/10 blocks for Stages 1–3, with learning rates 5e-5/5e-6/2e-6 and a 20%

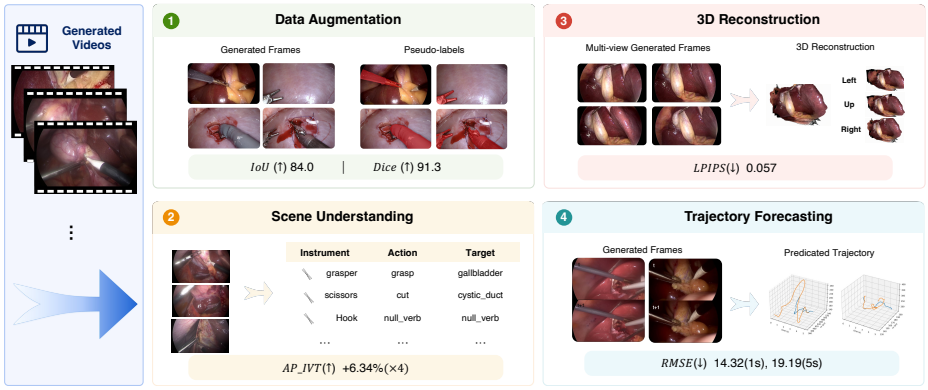

Fig. 2. Application-level utility of SurgGenesis. (1) Data augmentation. (2) Scene understanding. (3) 3D reconstruction. (4) Future-trajectory readout, evaluated by aligned frame-space RMSE; RMSE is multiplied by 100 for readability.

Cholec80 rehearsal ratio in Stages 2–3. Standard clips are generated at 832×480 in BF16 using 41 conditioning and 40 predicted frames. The temporal-horizon benchmark separately generates 81-frame future clips at 15 FPS, covering 5.33 s; the 1–5 s metrics use frame offsets 15/30/45/60/75, and a horizon is omitted when a clip does not provide sufficient coverage. Complementing the DiT as the forward state-transition model, an RT-DETR-based inverse dynamics model[24] recovers instrument actions from predicted frames and provides the trajectory readout. Controllability is judged separately by a CholecT50 recognizer trained only on real data. For the stagewise feature probe, we duplicate the final two WAN-VAE layers, extract their features, and apply a convolutional classifier; this lightweight head is fixed across stages to isolate representation changes. For segmentation, we append a lightweight decoder-side convolutional head that maps three channels to one mask channel and retrain the decoder/head, relying on the foundation model's surgical representation rather than a SOTA segmentation architecture. For the main IVT comparison, we reuse Rendezvous[20] solely as the triplet-recognition head; it is not used to generate segmentation masks. A separate fixed SAM-based evaluator measures pseudo-mask overlap between generated and real future frames. During downstream evaluation, the DiT remains frozen and only task heads are trained.

3.3 Ablation Study

We evaluate the three-stage strategy on generation fidelity and context-to-future consistency. The R3D-18 FVD-style score measures video fidelity, while frame-space RMSE compares generated frames with aligned real futures over 1–5s horizons. It is distinct from direct latent RMSE and the RT-DETR trajectory readout. Fig. 3 reports full stagewise curves: all adapted stages improve over native WAN in FVD-style fidelity, while Stage 2 and Stage 3 are closely matched

at 5s. Table 2 summarizes native-WAN-to-Stage-3 results. Fig. 3(a) and the Prediction Gap use native $[0, 1]$ RMSE; Fig. 2 and Table 2 report $100\times$ frame RMSE.

3.4 Downstream Tasks

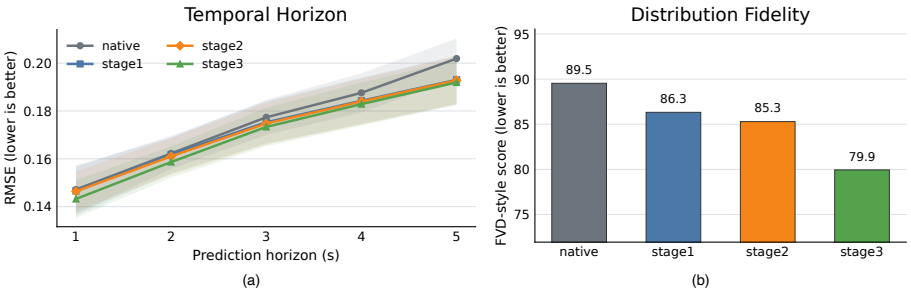

Fig. 3. Ablation and results. (a) Temporal-horizon frame RMSE on the native $[0, 1]$ scale. (b) Stage-wise R3D-18 FVD-style score.

We evaluate controlled surgical rollouts before asking whether they are useful downstream. Table 1 summarizes reference results for generation fidelity (LPIPS), rare-class augmentation (triplet mAP_{IVT}, following CholecT50[20]), and 3D reconstruction (LPIPS on VGGT4D[25] views). Because conditioning and protocols differ, the table provides context rather than a controlled head-to-head ranking. Distribution-level fidelity is reported in Fig. 3. We do not report FVD against unconditional generators such as Endora[16]: they synthesize from noise, whereas SurgGenesis rolls out a future from given frames and a structured prompt.

State-Transition Consistency: Predicted vs. Real Future. Visual fidelity does not establish that a model preserves the observed state. We therefore compare each rollout with its aligned real continuation. The Prediction Gap is computed in decoded frame space: GT latents give an upper-bound frame RMSE@5s of 0.0227, whereas Stage-3 predicted latents give 0.1888, for a gap of 0.1661. The dominant residual is therefore attributable to future prediction rather than VAE decoding. Direct latent RMSE is a separate diagnostic; the 1–5s curve reports aligned frame RMSE, and RT-DETR provides a complementary trajectory readout. This supplies state-transition evidence for the world-model formulation, not a claim about augmentation.

Text Controllability. An independent ConvNeXt-Tiny CholecT50 recognizer trained only on real data tests whether Stage-3 futures contain the IVT specified by the prompt. Across 432 clips and 24 minority IVT classes, it recovers the commanded triplet in **93.06%** of cases (instrument 97.45%, verb 90.97%, target 93.06%; five-fold range 87.5–96.6%), versus 6.25% for mismatched triplets.

Table 1. Generation on future frames. Reference results for fidelity (LPIPS↓), rare-class triplet augmentation (mAP_{IVT} ↑), and 3D reconstruction (LPIPS↓). Protocols differ; bold marks the best reported value, not a controlled ranking.

Method	Fidelity (LPIPS)↓	Triplet (mAP_{IVT})↑	3D (LPIPS)↓
SimDA[26]	0.565	32.7	0.089
VDM[27]	0.652	32.9	0.074
VidGPT[28]	0.563	35.8	0.089
EndoGen[29]	0.528	36.5	**0.047**
SurgGenesis	**0.516**	**36.9**	0.057

Table 2. Stagewise validation. The baseline is native WAN except for AP_{IVT}, which uses the real-only classifier. Native WAN failed under the 3D reconstruction pipeline, so its LPIPS is unavailable.

Downstream task	Metric	Baseline	Stage-3	Gain
Future trajectory	Frame RMSE@5s (×100) ↓	20.19	**19.19**	−4.94%
3D reconstruction	LPIPS ↓	Failed	**0.057**	–
SAM pseudo-mask probe	IoU / Dice ↑	0.2494 / 0.3117	**0.2615** / **0.3221**	+4.85/ + 3.34%
Video fidelity	R3D-18 FVD-style ↓	89.54	**79.94**	−10.72%
Long-tail recognition	AP_{IVT} ↑ (×4)	0.1717	**0.1823**	+6.34%

Mismatched prompts still recover shared instrument/verb components above 74%; target and full-triplet recovery are thus the more discriminative control tests. Together with state-transition consistency, this result supports context-grounded generation of the requested procedural state.

Application utility. Long-tail augmentation is an application-level test, not the definition of the world model. At matched 4× scale, SurgGenesis reaches 0.1823 AP_{IVT}, versus 0.1815 for Copy×4; rebalancing explains most of the gain, while synthesis adds modest action-conditioned variation. Under the non-identical protocols of Table 1, SurgGenesis is competitive in fidelity, triplet augmentation, and 3D reconstruction. The decoder-retrained segmentation head reaches 84.0 IoU / 91.3 Dice; Rendezvous is used only for IVT recognition. The fixed SAM diagnostic improves over native WAN (IoU 0.2615 vs. 0.2494; Dice 0.3221 vs. 0.3117), indicating useful action-specific variation without attributing the entire augmentation gain to synthesis.

Interpreting the evidence. FVD-style and LPIPS assess fidelity; the Prediction Gap and horizon-wise RMSE assess context-to-future consistency; and IVT recovery assesses prompt compliance. Segmentation, 3D reconstruction, and long-tail augmentation then test retained structure and downstream utility. The world-model claim rests on the first three levels together.

4 Conclusion

SurgGenesis is a text-controllable world model that rolls out futures from observed operative states and structured prompts. Fidelity, state-transition consistency, and IVT recovery assess its rollouts; downstream tasks test their utility.

References

1. Walimbe, S., Baby, B., Srivastav, V., Padoy, N.: Adaptation of multi-modal representation models for multi-task surgical computer vision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–33. Springer (2025)
2. Nwoye, C.I., Padoy, N.: Surgitrack: Fine-grained multi-class multi-tool tracking in surgical videos. Medical Image Analysis **101**, 103438 (2025)
3. Zia, A., Berniker, M., Nespolo, R., Zhang, X., 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). https://doi.org/10.48550/arXiv.2501.09209, https://arxiv.org/abs/2501.09209
4. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J., Li, S.: Mitral valve anatomy analysis using multimodal imaging data (2026). https://doi.org/10.5281/zenodo.19726755, https://doi.org/10.5281/zenodo.19726755, mICCAI 2026 Challenge
5. Yang, Y., Wang, Z.Y., Liu, Q., Sun, S., Wang, K., Chellappa, R., Zhou, Z., Yuille, A., Zhu, L., Zhang, Y.D., et al.: Medical world model: Generative simulation of tumor evolution for treatment planning. arXiv preprint arXiv:2506.02327 (2025)
6. 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 preprint arXiv:2512.23162 (2025)
7. 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 preprint arXiv:2511.01775 (2025)
8. Lin, H., Li, B., Wong, C.W., Rojas, J., Chu, X., Au, K.W.S.: World models for general surgical grasping. arXiv preprint arXiv:2405.17940 (2024)
9. 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)
10. Ding, T., Zou, Y., Chen, C., Shah, M., Tian, Y.: Clarity: Medical world model for guiding treatment decisions by modeling context-aware disease trajectories in latent space. arXiv preprint arXiv:2512.08029 (2025)
11. Xu, X., Li, Y., You, S., Bao, B.K.: Smrabooth: Subject and motion representation alignment for customized video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16130–16141 (2026)
12. Xu, X., Jia, G., Bao, B.K.: Disco-lora: Disentangled composition of content, style, and motion for multi-concept video customization. arXiv preprint arXiv:2606.26668 (2026)
13. 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)
14. Sun, W., You, X., Zheng, R., Yuan, Z., Li, X., He, L., Li, Q., Sun, L.: Bora: biomedical generalist video generation model. arXiv preprint arXiv:2407.08944 (2024)
15. Wang, H., Yang, Z., Zhang, H., Zhao, D., Wei, B., Xu, Y.: Feat: Full-dimensional efficient attention transformer for medical video generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 267–277. Springer (2025)

16. 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: International conference on medical image computing and computer-assisted intervention. pp. 230–240. Springer (2024)
17. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:2312.12429 (2023)
18. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
19. 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)
20. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
21. Nwoye, C.I., Elgohary, K., Srinivas, A., Zaid, F., Lavanchy, J.L., Padoy, N.: Cholectrack20: A multi-perspective tracking dataset for surgical tools. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8942–8952 (2025)
22. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International conference on medical image computing and computer-assisted intervention. pp. 431–441. Springer (2022)
23. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
24. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2024)
25. Hu, Y., Cheng, C., Yu, S., Guo, X., Wang, H.: Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. arXiv preprint arXiv:2511.19971 (2025)
26. Xing, Z., Dai, Q., Hu, H., Wu, Z., Jiang, Y.G.: Simda: Simple diffusion adapter for efficient video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7827–7839 (2024)
27. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
28. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
29. Liu, X., Liu, H., Wang, C., Liu, T., Yuan, Y.: Endogen: Conditional autoregressive endoscopic video generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 169–179. Springer (2025)