SurgHorizon: Long-Horizon Surgical World Modeling with Plausible Instrument-Tissue Dynamics

Wentao Pan[1,*] **Wuyang Li**[2,*] **Shengyuan Liu**[1]
Xinyu Liu[3] **Hengyu Liu**[1] **Yixuan Yuan**[1,†]
[1]The Chinese University of Hong Kong [2]EPFL [3]Imperial College London

Abstract

Scaling robot policy learning for autonomous surgery is challenging, as expert demonstrations are expensive and in vivo exploration poses substantial safety risks. Surgical world models address this by generating realistic, action-conditioned future frames from an initial observation, but existing methods exhibit two persistent failure modes: *spatial interaction incoherence*, where visible instrument contact fails to induce spatially consistent tissue deformation, and *temporal fidelity collapse*, where prediction errors compound across autoregressive rollouts and progressively corrupt visual quality. We present **SurgHorizon**, a surgical world model that mitigates both failures through two training recipes. *Deformation Consistency Regularization* extracts scene-point trajectories from training videos and enforces cross-frame coherence through latent contrastive learning, promoting plausible instrument-tissue dynamics. *Drift Adaptation Training* mitigates long-horizon drift by perturbing conditioning frames with online prediction residuals and photometric augmentations calibrated to long-horizon drift statistics, sustaining visual fidelity over extended rollouts. To enable rigorous evaluation, we further introduce **SurgWorld-Bench**, featuring diverse procedure types, long-range rollouts, and decoupled metrics for instrument-motion accuracy and tissue-response fidelity. Extensive experiments show that SurgHorizon consistently outperforms state-of-the-art methods across visual quality, temporal consistency, and interaction fidelity, with gains widening as the prediction horizon grows.

1 Introduction

Autonomous surgery holds transformative potential to reduce surgeon workload, improve the quality of operative care, and expand access to high-quality surgical treatment [32, 12]. However, scaling robot policy learning through real-world interaction remains difficult, as expert demonstrations are expensive to acquire and in vivo exploration poses substantial safety risks. To address this, Surgical World Models (SWMs) [20, 10, 31] offer a promising pathway for learning in simulated surgical environments (Fig. 1a). Conditioned on an initial observation and planned instrument actions, such as tool trajectories, an SWM synthesizes future intraoperative frames that capture instrument motion and tissue responses, enabling visual simulation of surgical interactions (Fig. 1b).[1]

Recent SWM works mainly focus on two complementary directions. To improve visual realism, several works [20, 37, 31] transfer foundational video DiTs [41, 40] to the surgical domain, producing realistic endoscopic videos under instruction-based control. To improve action controllability, the control signal has progressed from coarse textual descriptions [20, 37] to fine-grained spatial specifi-

[*]Equal contribution. [†]Corresponding author.

[1]Unlike generic-domain world models [15, 36, 19], SWMs must capture plausible instrument-tissue dynamics. This requires modeling contact-driven tissue deformation under challenging endoscopic imaging conditions, including low-light or non-uniform illumination, specular reflections, smoke, and occlusions.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

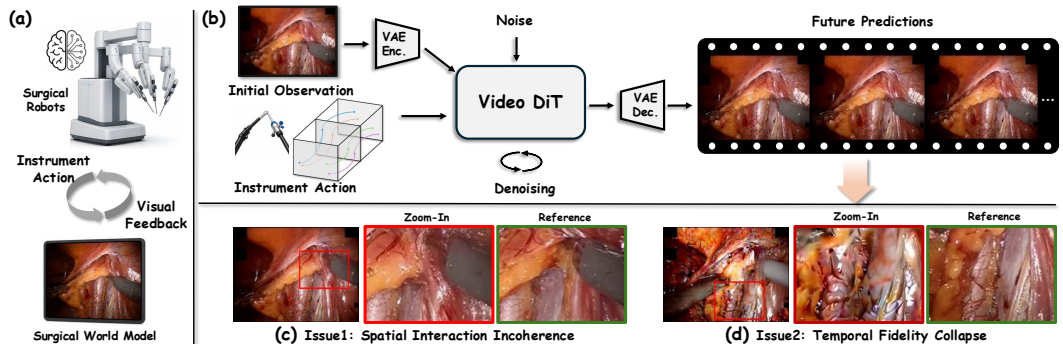

Figure 1: **Surgical world modeling.** (a) Simulating intraoperative dynamics for policy learning. (b) Predicting future frames from an initial frame and instrument actions. **Two key challenges:** (c) Tissue fails to deform under instrument contact. (d) Visual quality degrades over long-horizon rollouts.

cations [3, 10, 31], allowing more accurate instrument motion control. These advances are essential for building high-fidelity visual simulators to support controllable policy learning and evaluation.

Despite progress, existing SWMs still face two key challenges. **(i) Spatial Interaction Incoherence.** Existing works directly optimize pixel distribution without imposing explicit cross-frame motion constraints, often causing spatially inconsistent tissue responses under visible instrument contact (Fig. 1c). While some methods [10, 31] incorporate geometric cues such as depth priors to mitigate this issue, tool-induced tissue deformation in surgical scenes is highly localized and morphologically complex, requiring fine-grained motion correspondence that global geometric priors cannot provide. **(ii) Temporal Fidelity Collapse.** Existing SWMs suffer from severe degradation in long-horizon simulation. Over rollouts spanning tens of seconds, generated frames progressively exhibit overexposure, color drift, and structural corruption (Fig. 1d), as prediction errors propagate through autoregressive rollouts [22, 13]. Complex tissue textures and challenging illumination further amplify this degradation, limiting the effectiveness of general-domain solutions [26, 39].

To address these issues, we present **SurgHorizon**, a surgical world model that achieves plausible instrument-tissue dynamics over long horizons, building upon two complementary training strategies. *(i) Deformation Consistency Regularization* extracts dense scene-point trajectories that jointly cover instruments and tissue, and enforces contrastive coherence in the low-dimensional manifold, improving instrument-tissue motion consistency. *(ii) Drift Adaptation Training* mitigates long-horizon drift by perturbing conditioning frames with online prediction residuals and rollout-calibrated photometric augmentations, sustaining visual fidelity over extended rollouts. We further propose **SurgWorld-Bench**, a benchmark designed for long-range surgical simulation across diverse procedure types, featuring hierarchical evaluation of instrument-motion accuracy and tissue-response fidelity. Extensive experiments show that SurgHorizon consistently outperforms state-of-the-art methods, with gains increasing as the horizon lengthens. Our contributions are as follows:

- We present SurgHorizon, a surgical world model that mitigates spatial interaction incoherence and temporal fidelity collapse through two dedicated training recipes, enabling long-horizon surgical simulation with plausible instrument-tissue interactions.
- We introduce Deformation Consistency Regularization, which enforces cross-frame coherence along scene-point trajectories, suppressing incoherent instrument-tissue dynamics.
- We introduce Drift Adaptation Training, which perturbs training conditions with online prediction residuals and photometric degradations, enabling longer-horizon stable rollouts.
- We present SurgWorld-Bench, a multi-horizon multi-procedure benchmark with decoupled metrics for instrument-motion accuracy and tissue-response fidelity.

2 Related Work

2.1 Surgical World Models

Classical world models learn compact latent dynamics for policy optimization in simulated environments [15–17], while recent video diffusion models demonstrate that pixel-space generation itself

can approximate a simulator [5, 36, 19, 25, 28]. In surgery, existing world models formulate future prediction as pixel-wise video generation conditioned on varying action representations. One line of work [20, 37, 31] fine-tunes foundation video DiTs [41, 40] on surgical footage, enabling surgical world modeling with improved realism. In parallel, the granularity of action control has advanced from coarse textual descriptions [20, 37] to fine-grained spatial signals such as panoptic segmentation maps [3], optical flow [10], and affordance masks combined with tool-tip trajectories [31]. Recent efforts further incorporate depth priors as auxiliary inputs [10] or supervision [31] to improve geometric plausibility, yet cross-frame motion correspondence remains underexplored, leading to spatial interaction incoherence. Moreover, existing methods are evaluated on short clips and suffer from temporal fidelity collapse over long-horizon predictions. SurgHorizon targets both gaps with two dedicated training recipes: Deformation Consistency Regularization to enforce cross-frame interaction coherence, and Drift Adaptation Training to sustain fidelity over long horizons.

2.2 Long-Horizon World Modeling

Extending video generation models to long-horizon prediction typically relies on autoregressive conditioning, where each new clip is generated from preceding frames. This introduces a train-inference distribution gap [29, 33]: the model sees clean ground-truth contexts during training yet conditions on its own imperfect predictions at inference, causing errors to accumulate over time. Existing remedies fall into two categories. The first reduces dependency on preceding frames through modified noise schedules [7, 35], anchor frames [21, 39], or anti-drifting sampling [42], alleviating but not eliminating drift. The second exposes the model to degraded conditioning during training to build robustness. Self-Forcing-style methods [22, 13] unroll multi-step autoregressive rollouts within each iteration, but require expensive bidirectional-to-causal distillation. Error recycling [26, 39] offers a lighter alternative by re-injecting previously collected prediction errors into clean inputs, but the stored errors originate from different samples or outdated checkpoints, poorly approximating the degradation that the current model would actually produce. Our Drift Adaptation Training sidesteps these limitations: a single forward step probes sample-specific degradation while keeping the simulated drift synchronized with the current model, and photometric augmentations further simulate pixel-level artifacts that latent-space error injection alone overlooks.

3 SurgHorizon

Problem Formulation. Given an observed frame $\mathbf{s}_0 \in \mathbb{R}^{H\times W\times 3}$ and an action sequence $\mathbf{a}_{1:T}$, SurgHorizon predicts future frames $\hat{\mathbf{s}}_{1:T}$ that capture both instrument motion and the induced tissue deformation. For long-horizon simulation beyond a single clip, generated clips are chained autoregressively, where the tail frames of each generated clip condition the next.

Overview. SurgHorizon builds on a pretrained latent video DiT [38]. Let $\mathbf{z}_1 = \mathcal{E}(\mathbf{s}_{1:T}) \in \mathbb{R}^{T_z\times H_z\times W_z\times C}$ denote the target latent encoded by a pretrained VAE encoder $\mathcal{E}$, and let $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the noise sample. Flow matching constructs the interpolant $\mathbf{z}_\tau = \tau\mathbf{z}_1 + (1-\tau)\mathbf{z}_0$ at flow time $\tau \sim \mathcal{U}[0,1]$ and trains the DiT to regress the velocity field $v_\theta(\mathbf{z}_\tau, \tau)$. To condition generation on instrument actions, each action $\mathbf{a}_t$ is represented as a set of instrument keypoint coordinates, forming trajectories $\{\mathbf{p}^k_{1:T}\}_{k=1}^K$ that specify instrument positions over time. Latent Action Encoding (Sec. 3.1) converts these trajectories into an action latent $\mathbf{M}$ by replicating first-frame latent features at each trajectory's target positions. The action latent $\mathbf{M}$ is concatenated with $\mathbf{z}_\tau$ along the channel dimension, while $\mathbf{z}_{\text{first}} = \mathcal{E}(\mathbf{s}_0)$ is projected through a reference convolution into tokens prepended to the DiT input sequence. Text prompts are encoded and injected via cross-attention. We denote all conditioning collectively as $\mathbf{c}$. The model is trained by minimizing

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau,\,\mathbf{z}_0}\left[\left\|v_\theta(\mathbf{z}_\tau, \mathbf{c}, \tau) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2\right]. \tag{1}$$

This pipeline produces visually plausible short clips, yet two limitations remain. The flow-matching objective does not enforce cross-frame motion coherence, leading to spatial interaction incoherence in instrument-tissue dynamics. Autoregressive rollout further exposes the model to its own imperfect predictions, causing temporal fidelity collapse over long horizons. SurgHorizon introduces two training strategies: **Deformation Consistency Regularization** (DCR, Sec. 3.2), which enforces cross-frame motion coherence along tracked scene-point trajectories, and **Drift Adaptation Training** (DAT, Sec. 3.3), which closes the train-inference conditioning gap through online drift simulation.

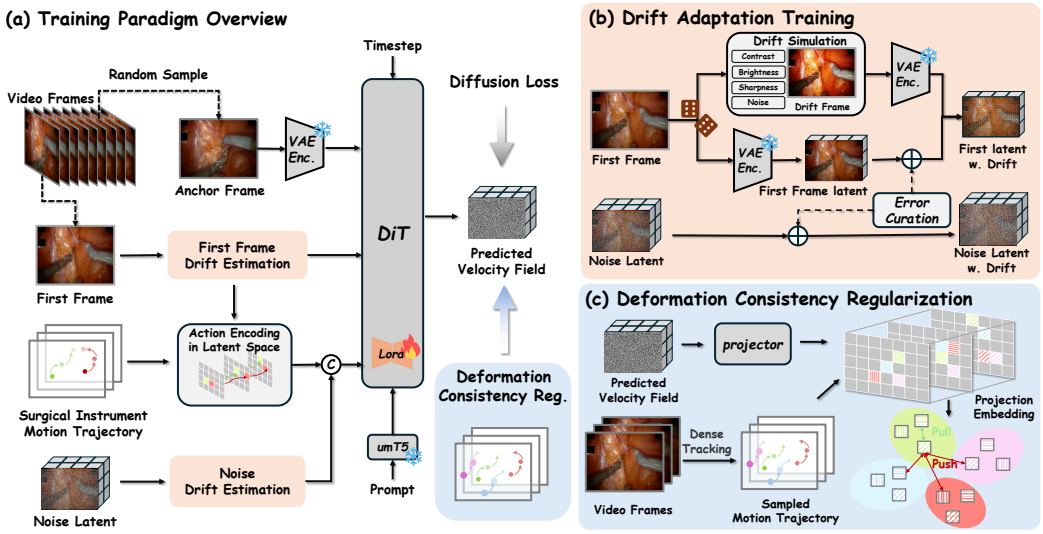

Figure 2: **Overview of SurgHorizon.** (a) Training paradigm, consisting of two training recipes: Deformation Consistency Regularization (DCR) and Drift Adaptation Training (DAT). (b) DAT simulates inference-time conditioning drift via two complementary proxies. (c) DCR enforces cross-frame motion coherence via contrastive learning on tracked scene-point trajectories.

3.1 Latent Action Encoding

Prior surgical world models encode instrument actions through segmentation masks [3], optical flow [10], or trajectory videos [31], requiring auxiliary models that increase computational overhead. SurgHorizon instead converts instrument keypoint trajectories into an action latent by replicating first-frame latent features along each trajectory [11]. Concretely, given the initial-frame latent $\mathbf{z}_{\text{first}} = \mathcal{E}(\mathbf{s}_0) \in \mathbb{R}^{H_z \times W_z \times C}$ and pixel-space trajectories $\{\mathbf{p}^k_{1:T}\}_{k=1}^K$, each trajectory is mapped onto the spatiotemporal latent grid by downscaling spatial coordinates by the VAE spatial factor f_s and grouping every f_t consecutive frames into one latent frame, yielding latent-space trajectories $\{\hat{\mathbf{p}}^k_{1:T_z}\}_{k=1}^K$ with $T_z = T/f_t$. The latent feature at each keypoint's initial position is then replicated to its target position in every subsequent latent frame, producing the action latent $\mathbf{M} \in \mathbb{R}^{T_z \times H_z \times W_z \times C}$:

$$\mathbf{M}\big[\hat{t},\, \hat{\mathbf{p}}^k_{\hat{t}}\big] = \mathbf{z}_{\text{first}}\big[\hat{\mathbf{p}}^k_1\big], \quad \forall k,\, \hat{t} > 0, \tag{2}$$

with all other entries set to zero. The resulting $\mathbf{M}$ supplies spatially grounded action guidance without introducing any auxiliary modules.

3.2 Deformation Consistency Regularization

Surgical world modeling demands motion coherence of instrument-tissue dynamics across frames, yet standard generative training solely models pixel distributions and provides no mechanism to enforce such correspondence. As a result, generated sequences can exhibit interaction incoherence: an instrument contacts tissue, yet the surrounding region fails to produce a visually coherent deformation response. To address this, we introduce Deformation Consistency Regularization (DCR), which imposes a motion-coherence constraint on the predicted video latent, as illustrated in Fig. 2c.

Keypoint trajectories offer a unified motion description for both rigid instruments and deformable tissue, making them a natural supervision signal for cross-frame coherence. We treat such trajectories as cross-frame correspondence anchors and enforce that the predicted latent assigns a consistent representation to each tracked point over time. Specifically, these anchors are sampled across both instrument and tissue regions, with contrastive positive pairs formed strictly along the temporal trajectory of the *same* physical point to preserve both tool motion and tissue deformation.

We employ an off-the-shelf point tracker [24] to extract dense grid trajectories $\{(\mathbf{q}^j_t,\, o^j_t)\}_{t=1}^T$ for $j = 1, \ldots, J$ from each training video (details in Appendix C), where $\mathbf{q}^j_t \in \mathbb{R}^2$ denotes the image-space location of point j at frame t and $o^j_t \in [0, 1]$ its visibility. At each training step, given a sampled

flow time $\tau \sim \mathcal{U}[0,1]$ with noisy latent $\mathbf{z}_\tau$, we obtain the predicted clean latent via a single Euler step,

$$\hat{\mathbf{z}}_1 = \mathbf{z}_\tau + (1-\tau)\cdot v_\theta(\mathbf{z}_\tau,\, \mathbf{c},\, \tau)\,, \tag{3}$$

and extract each tracked point's feature from the corresponding frame slice $\hat{\mathbf{z}}_{1,t}$ at latent-grid coordinate $\mathbf{q}_t^j/f_s$, projecting it through a trainable two-layer MLP ϕ into a d-dimensional embedding space,

$$\mathbf{f}_t^j = \phi\Big(\mathrm{sample}\Big(\hat{\mathbf{z}}_{1,t},\, \mathbf{q}_t^j/f_s\Big)\Big)\,, \tag{4}$$

yielding an ℓ_2-normalized descriptor $\mathbf{f}_t^j \in \mathbb{R}^d$ at every visible frame $t \in \mathcal{O}^j = \{t : o_t^j > o_{\min}\}$.

For each trajectory j, a prototype $\bar{\mathbf{f}}^j = \frac{1}{|\mathcal{O}^j|}\sum_{t\in\mathcal{O}^j}\mathbf{f}_t^j$ aggregates descriptors over visible frames. An InfoNCE loss [9, 18] then pulls each descriptor toward its own prototype while pushing it away from the prototypes of the remaining $J-1$ trajectories:

$$\mathcal{L}_{\text{DCR}} = -\frac{1}{|\mathcal{Q}|}\sum_{(j,t)\,\in\,\mathcal{Q}} \log \frac{\exp\Big(\mathbf{f}_t^j\cdot\bar{\mathbf{f}}^j/\gamma\Big)}{\displaystyle\sum_{j'=1}^{J}\exp\Big(\mathbf{f}_t^j\cdot\bar{\mathbf{f}}^{j'}/\gamma\Big)}, \tag{5}$$

with $\mathcal{Q} = \{(j,t) : t \in \mathcal{O}^j\}$ and temperature γ. The positive term anchors each tracked point to a stable representation across time, while the negatives prevent trivial collapse. Together with the flow-matching objective, DCR enforces cross-frame latent consistency aligned with instrument motion and tissue deformation within each generated clip.

3.3 Drift Adaptation Training

When generation extends beyond a single clip through autoregressive conditioning, a train-inference distribution shift emerges: the model trains on clean ground-truth frames yet conditions on its own imperfect outputs at inference. This mismatch compounds across clips and eventually dominates long-horizon failure. The key idea of DAT is to deliberately corrupt conditioning frames during training, forcing the model to recover clean predictions from drifted inputs and thereby building robustness to imperfect inference-time conditions. DAT replaces clean conditioning with corrupted surrogates through two complementary proxies (Fig. 2b), one model-intrinsic and one model-agnostic.

Model-intrinsic proxy: online prediction residuals. An ideal training proxy would expose the model to perturbations drawn from the same error distribution it will encounter during rollout. We achieve this by extracting the model's own prediction error from the training forward pass and injecting it back as a conditioning perturbation.

Specifically, we sample an independent flow time $\tau' \sim \mathcal{U}[0,1]$, construct $\mathbf{z}_{\tau'} = \tau'\mathbf{z}_1 + (1-\tau')\mathbf{z}_0'$ with a fresh noise sample $\mathbf{z}_0' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and obtain the model's own prediction of the clean video latent via a single no-gradient Euler step,

$$\hat{\mathbf{z}}_1' = \mathbf{z}_{\tau'} + (1-\tau')\cdot \mathrm{sg}[v_\theta(\mathbf{z}_{\tau'},\, \mathbf{c},\, \tau')]\,, \tag{6}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient and $\mathbf{c}$ uses the original clean conditioning. The first temporal slice $\hat{\mathbf{z}}_{1,0}'$ is the model's prediction of the first frame. Its deviation from the ground-truth conditioning latent yields the drift residual,

$$\boldsymbol{\delta} = \hat{\mathbf{z}}_{1,0}' - \mathbf{z}_{\text{first}}, \tag{7}$$

which is injected back onto the clean conditioning to simulate corrupted inference-time inputs,

$$\tilde{\mathbf{z}}_{\text{first}} = \mathbf{z}_{\text{first}} + \boldsymbol{\delta}. \tag{8}$$

In addition, the full prediction error is injected into the noisy latent $\tilde{\mathbf{z}}_\tau = \mathbf{z}_\tau + (\hat{\mathbf{z}}_1' - \mathbf{z}_1)$, so that the denoising input jointly reflects the drift the model will encounter at inference. Since $\boldsymbol{\delta}$ originates from the model's own imperfect prediction, it naturally mirrors the error characteristics the model will produce during autoregressive rollout. Sampling τ' and $\mathbf{z}_0'$ independently of the main loss decouples the corruption severity from the training noise level, increasing the diversity of simulated drift. We detach $\boldsymbol{\delta}$ from the computational graph but recompute it at every iteration, so the proxy co-evolves with v_θ without affecting its gradient.

Model-agnostic proxy: photometric augmentations. As training progresses, the probe residual $\boldsymbol{\delta}$ shrinks with improving model accuracy. Yet repeated decode-encode cycles during autoregressive rollout still introduce persistent photometric degradations, including contrast drift, color skew, texture blur, and specular blowout. To quantify this effect, we measure photometric statistics over long-horizon rollouts and find that four key dimensions (brightness, contrast, sharpness, and noise level) diverge progressively from ground truth as the horizon grows (Fig. 3). Motivated by this observation, we randomly apply four photometric perturbations to the conditioning frame: exposure shift, contrast jitter, sharpness scaling, and noise adjustment, each with randomly sampled strength calibrated to match observed rollout artifacts. The corrupted conditioning latent is then obtained as

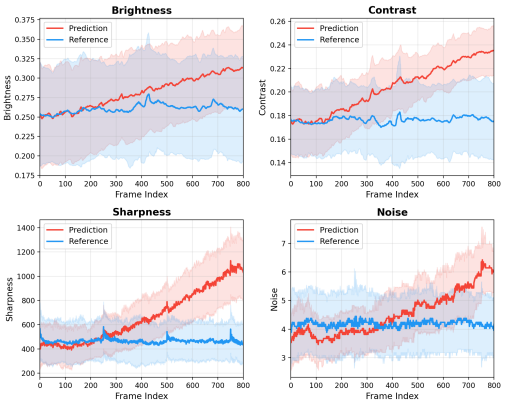

Figure 3: Main photometric statistics over long-horizon predictions.

$$\tilde{\mathbf{z}}_{\text{first}} = \mathcal{E}(\mathcal{T}_{\text{img}}(\mathbf{s}_{\text{first}}))\,, \tag{9}$$

where $\mathcal{T}_{\text{img}}$ denotes the sampled augmentation. Since $\mathcal{T}_{\text{img}}$ is independent of v_θ, this proxy provides a persistent source of corrupted conditioning regardless of the model's current prediction quality.

Training objective. At each training step, we randomly select one of three conditioning modes: the model-intrinsic proxy (Eq. 8), the model-agnostic proxy (Eq. 9), or the clean latent $\mathbf{z}_{\text{first}}$ to preserve accuracy on uncorrupted inputs, with probabilities p_{probe}, p_{aug}, and p_{clean} respectively. The target remains the clean ground-truth clip regardless of the conditioning mode. The flow-matching loss under the selected conditioning is

$$\mathcal{L}_{\text{DAT}} = \mathbb{E}_{\tau,\,\mathbf{z}_0}\left[\left\|v_\theta(\tilde{\mathbf{z}}_\tau,\,\tilde{\mathbf{c}},\,\tau) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2\right], \tag{10}$$

where $\tilde{\mathbf{c}}$ denotes the conditioning with $\mathbf{z}_{\text{first}}$ replaced by the selected corrupted variant $\tilde{\mathbf{z}}_{\text{first}}$ and supplemented with an anchor frame $\mathbf{s}_{\text{anchor}}$ fixed to the first frame throughout inference as a clean visual reference, and $\tilde{\mathbf{z}}_\tau$ equals $\mathbf{z}_\tau + (\hat{\mathbf{z}}'_1 - \mathbf{z}_1)$ under the model-intrinsic proxy and reduces to the unmodified $\mathbf{z}_\tau$ under the other two modes.

Remark. Recent error-recycling methods [26, 39] also target the train-inference gap, but draw corrupted conditioning from heterogeneous sources such as different samples or earlier checkpoints that may not reflect the current model's error distribution. DAT instead computes the residual online from the same sample in each forward pass, so the simulated drift co-evolves with the model. The photometric proxy further covers pixel-level artifacts that latent-space residuals alone cannot capture.

Combining the two training recipes, the full objective of SurgHorizon is

$$\mathcal{L} = \mathcal{L}_{\text{DAT}} + \lambda_{\text{DCR}}\,\mathcal{L}_{\text{DCR}}, \tag{11}$$

where $\mathcal{L}_{\text{DAT}}$ drives temporal fidelity under conditioning drift and $\mathcal{L}_{\text{DCR}}$ enforces cross-frame motion coherence within each generated clip.

4 SurgWorld-Bench

Existing evaluations of surgical world models share three limitations: coverage of only a single procedure type, rollout lengths rarely exceeding 80 frames [20, 3, 10, 31], and reliance on holistic metrics that conflate instrument motion with tissue response. We introduce **SurgWorld-Bench**, a surgical world-model benchmark to *jointly* address all three through (a) *procedural diversity*, (b) *long-horizon evaluation* up to $10\times$ the typical 80-frame rollout, and (c) a *decoupled evaluation protocol* that separately probes instrument motion and tissue response (Fig. 4).

Procedural diversity. SurgWorld-Bench draws from three public datasets covering three procedures (Fig. 4a): GraSP [2] for radical prostatectomy, SurgToolLoc22 [44] for porcine laparoscopic procedures, and CholecTrack20 [30] for laparoscopic cholecystectomy, encompassing diverse subjects, heterogeneous instruments, and varied tissue deformation dynamics.

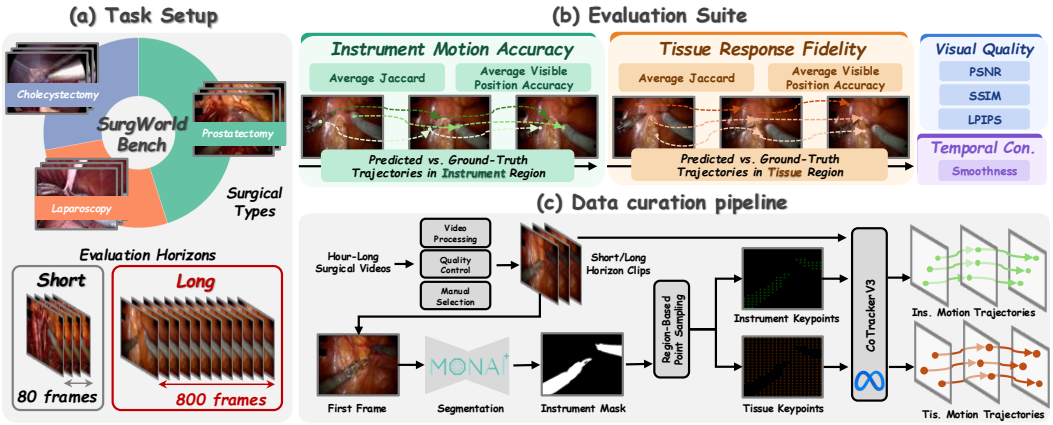

Figure 4: **Overview of SurgWorld-Bench.** (a) Task setup with three procedure types and two evaluation horizons. (b) Evaluation suite with decoupled instrument and tissue trajectory metrics. (c) Data curation pipeline for extracting region-specific motion trajectories.

Task and evaluation horizons. Given an initial frame $\mathbf{s}_0$ and an instrument trajectory $\mathbf{a}_{1:T}$, the model predicts the future state sequence $\hat{\mathbf{s}}_{1:T}$. The benchmark operates at two evaluation horizons (Fig. 4a): a short horizon of 80 frames over 111 evaluation clips for comparability with prior work, and a long horizon of 800 frames over 21 evaluation clips to probe long-horizon stability.

Evaluation protocol. Holistic metrics (e.g., PSNR) alone cannot distinguish whether errors originate from instrument motion or tissue response. SurgWorld-Bench therefore assesses predictions along four complementary dimensions (Fig. 4b), where the first two retain the standard global view and the latter two supply region-decoupled diagnosis. **(i) Visual quality**: per-frame PSNR, SSIM, and LPIPS [43]. **(ii) Temporal consistency**: the Motion Smoothness Score [23], which captures temporal coherence. For **(iii) Instrument-motion accuracy** and **(iv) Tissue-response fidelity**, initial grid points are partitioned into instrument and tissue regions via first-frame instrument segmentation [6], and a point tracker [24] tracks these points in both predicted and ground-truth sequences (Fig. 4c). Two tracking metrics are reported for each region: Average Jaccard (AJ) and Average Visible Position Accuracy ($\delta^{\text{vis}}_{\text{avg}}$) [14]. The instrument-region scores measure whether predicted instruments follow the prescribed action $\mathbf{a}_{1:T}$, while the tissue-region scores measure the consistency of predicted tissue dynamics with the ground-truth observation. Details are provided in Appendix A.

5 Experiments

5.1 Setup

SurgHorizon builds on Wan2.1-1.3B [1] and is trained for 100K iterations at 512×320 resolution with a clip length of 81 frames. Baselines fall into two groups: general-purpose video generation models conditioned on text-described actions (SVD [4], CogVideoX1.5-5B [41], HunyuanVideo-1.5 [40], Wan2.2-TI2V-5B [38]), and controllable video generation models conditioned on dense action signals such as optical flow or point trajectories (Wan2.1-Control [1], SurgSora [10]). All methods are evaluated on SurgWorld-Bench under short-horizon and long-horizon settings. Dataset sources, training configurations, and additional implementation details are provided in Appendix C.

5.2 Comparison with State-of-the-Art Methods

Quantitative results. Table 1 reports the main comparison on SurgWorld-Bench. SurgHorizon outperforms all baselines across every evaluation dimension, and its advantage widens as the prediction horizon increases. For short-horizon world modeling, general-purpose video generation models produce visually realistic frames yet lack fine-grained controllability over instrument motion, resulting in instrument motions that deviate from the specified actions. SurgSora, designed specifically for surgical video modeling, exhibits stronger instrument control but is constrained by its backbone capacity, trailing SurgHorizon in tissue-response fidelity and visual quality. The cross-frame motion

Table 1: **Comparison with state-of-the-art methods on SurgWorld-Bench** across instrument-motion accuracy, tissue-response fidelity, visual quality, and temporal consistency at short-horizon and long-horizon settings, with best results per column in **bold**.

	Method	Instrument Motion		Tissue Response		Visual Quality			Temporal Consistency
		AJ ↑	δ_{avg}^{vis} ↑	AJ ↑	δ_{avg}^{vis} ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Smoothness ↑
Short	SVD [4]	3.62	7.19	8.69	16.22	13.80	0.42	0.48	96.61
	CogVideoX1.5-5B [41]	4.34	8.24	15.21	26.46	13.14	0.42	0.46	96.91
	HunyuanVideo-1.5 [40]	4.80	9.89	18.93	30.91	16.35	0.51	0.36	**99.14**
	Wan2.2-TI2V-5B [38]	5.83	10.40	19.93	32.32	15.63	0.51	0.40	98.77
	Wan2.1-Control [1]	39.34	52.08	36.19	49.51	19.08	0.62	0.26	99.03
	SurgSora [10]	19.41	27.00	19.51	30.51	15.12	0.43	0.49	98.50
	SurgHorizon (Ours)	**45.76**	**58.43**	**41.68**	**55.29**	**19.85**	**0.65**	**0.23**	**99.14**
Long	SVD [4]	1.61	2.46	3.55	5.54	7.99	0.23	0.76	96.26
	CogVideoX1.5-5B [41]	2.69	4.06	7.65	11.41	10.44	0.39	0.67	96.75
	HunyuanVideo-1.5 [40]	1.56	3.73	5.63	11.76	14.24	0.41	0.57	99.07
	Wan2.2-TI2V-5B [38]	1.94	3.15	5.91	9.57	11.87	0.39	0.65	98.74
	Wan2.1-Control [1]	18.92	28.30	19.91	31.90	16.35	0.50	0.42	99.51
	SurgSora [10]	5.36	7.57	6.08	9.37	8.95	0.28	0.79	98.55
	SurgHorizon (Ours)	**27.13**	**39.61**	**24.15**	**38.19**	**17.36**	**0.55**	**0.37**	**99.53**

supervision of DCR enables SurgHorizon to achieve more accurate instrument control and more realistic tissue responses while maintaining superior visual quality. Long-horizon world modeling poses a greater challenge, as all methods suffer degradation in visual quality, instrument-control accuracy, and tissue-response realism. By simulating conditioning drift during training, DAT improves rollout stability, maintaining SurgHorizon's clear lead over the baselines.

Qualitative comparison. Figure 5 presents a visual comparison of SurgHorizon against SurgSora and Wan2.1-Control under both short- and long-horizon settings. In the short-horizon setting, SurgSora exhibits inaccurate instrument motion accompanied by noticeable visual artifacts. Wan2.1-Control reproduces instrument motion faithfully, yet the surrounding tissue remains largely static, missing the deformations expected upon instrument contact. SurgHorizon captures both accurate instrument motion and the resulting tissue response, with highlighted regions revealing realistic deformation around the interaction point. Over longer horizons, SurgSora nearly collapses, while Wan2.1-Control accumulates color drift and structural blur. By contrast, SurgHorizon maintains realistic appearance, faithful instrument geometry, and plausible tissue responses.

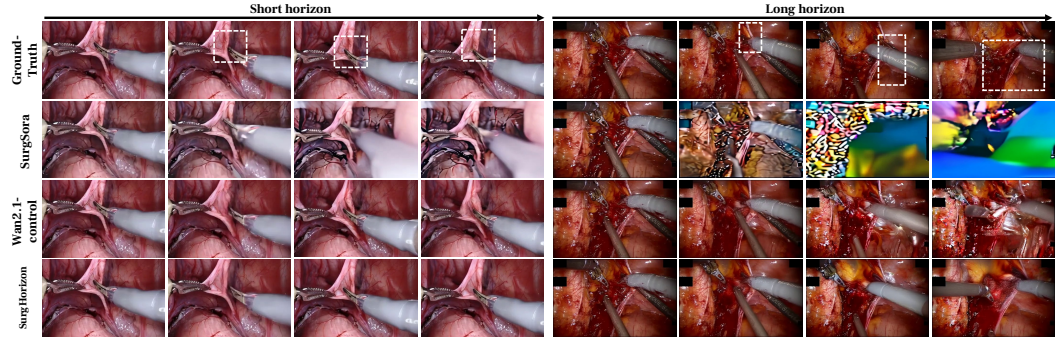

Figure 5: **Qualitative comparison with state-of-the-art methods.** Left: short-horizon predictions at frame indices 1, 20, 40, and 60. Right: long-horizon predictions at frames 1, 100, 500, and 800.

5.3 Ablation Studies

Component ablation. Table 2 ablates the contributions of DCR and DAT under long-horizon prediction. DCR alone improves instrument-motion and tissue-deformation accuracy through cross-frame motion regularization. DAT alone yields larger gains by simulating conditioning drift during

Table 2: **Component ablation** on instrument and tissue accuracy (long-horizon).

	Ins.		Tis.	
Method	AJ	δ^{vis}_{avg}	AJ	δ^{vis}_{avg}
Baseline	19.45	29.09	20.00	32.20
+ DCR only	20.57	31.10	21.18	33.67
+ DAT only	26.54	38.83	24.10	37.96
+ DCR + DAT (full)	**27.13**	**39.61**	**24.15**	**38.19**

Table 3: **Effect of drift simulation strategies** on accuracy (long-horizon).

	Ins.		Tis.	
Method	AJ	δ^{vis}_{avg}	AJ	δ^{vis}_{avg}
Clean only	22.11	33.64	21.87	35.16
Error recycling [26]	24.15	35.96	23.35	36.76
DAT – latent	26.09	38.55	24.09	37.74
DAT – latent + image	**27.13**	**39.61**	**24.15**	**38.19**

Table 4: **Effect of DCR spatial region** on accuracy (short-horizon).

	Ins.		Tis.	
Method	AJ	δ^{vis}_{avg}	AJ	δ^{vis}_{avg}
Baseline	40.42	53.13	36.07	49.43
DCR – Instrument only	44.91	57.62	38.39	52.29
DCR – Tissue only	44.42	57.21	41.23	54.90
DCR – All (default)	**45.76**	**58.43**	**41.68**	**55.29**

Table 5: **Effect of DCR trajectory count** on accuracy (short-horizon).

	Ins.		Tis.	
Method	AJ	δ^{vis}_{avg}	AJ	δ^{vis}_{avg}
Baseline	40.42	53.13	36.07	49.43
DCR – 32	44.65	57.25	41.07	54.68
DCR – 64 (default)	**45.76**	**58.43**	41.68	55.29
DCR – 128	45.74	58.37	**41.70**	**55.37**

training, targeting the autoregressive distribution gap that dominates long-horizon failure. Combining both achieves the best results across all metrics, indicating that spatial coherence from DCR and drift robustness from DAT address complementary aspects of long-horizon degradation.

Drift simulation strategies. Table 3 compares three drift simulation strategies sharing DCR and the anchor frame: clean-only training, error recycling from [26], and our full DAT design. Clean training degrades rapidly under accumulated rollout drift. Error recycling [26] reuses prediction residuals from other samples and earlier checkpoints, offering limited adaptation. The online self-prediction strategy in DAT recomputes per-sample residuals under current model weights, producing drift signals well aligned with the ongoing training state. Coupling this with image-space photometric perturbations captures pixel-level drift, yielding peak long-horizon performance.

DCR configuration design. To examine how supervision regions affect DCR, constraints are applied to instrument, tissue, or full-frame regions separately. As shown in Table 4, supervising tissue regions improves tissue-response fidelity over instrument-only constraints, highlighting the value of explicit tissue correspondence in capturing dynamic deformation. Full-frame supervision achieves the best results across all metrics, supporting joint regularization of tool and tissue regions. We also vary the number of sampled trajectories in Table 5. Increasing from 32 to 64 trajectories yields clear gains as denser sampling covers richer interaction dynamics, whereas scaling to 128 saturates performance. We thus set 64 as default to balance supervision quality and computational cost.

6 Limitations

Despite stable long-horizon rollouts, SurgHorizon relies on 2D visual point trajectories as control signals rather than true robot kinematics or physical interaction forces. Furthermore, simulating extreme non-rigid deformations during rapid tool motions and preserving delicate structures like fine sutures remain challenging. Detailed analysis is provided in Appendix D.

7 Conclusion

This paper presents SurgHorizon, a surgical world model developed for stable long-horizon simulation with realistic instrument-tissue dynamics. To resolve spatial interaction incoherence, Deformation Consistency Regularization aligns cross-frame latent trajectories to promote plausible instrument-tissue dynamics. To curb cumulative drift during extended rollouts, Drift Adaptation Training closes the train-inference exposure gap via online residual perturbation and photometric calibration. Evaluations on proposed SurgWorld-Bench confirm consistent gains over competitive baselines across varying horizons, providing a dependable foundation for interactive surgical policy learning.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by National Natural Science Foundation of China (No. 62522124), Hong Kong Research Grants Council General Research Fund 14206725, Hong Kong Innovation and Technology Commission Innovation and Technology Fund PRP/082/24FX, and CRS_CityU103/25.

The authors declare no competing interests.

References

[1] *Alibaba PAI* . Wan2.1-Fun-1.3B-Control. 2025. Hugging Face model repository.

[2] *Ayobi Nicolás, Rodríguez Santiago, Pérez Alejandra, Hernández Isabela, Aparicio Nicolás, Dessevres Eugénie, Peña Sebastián, Santander Jessica, Caicedo Juan Ignacio, Fernández Nicolás, others* . Pixel-wise recognition for holistic surgical scene understanding // Medical Image Analysis. 2025. 103726.

[3] *Biagini Diego, Navab Nassir, Farshad Azade*. Hierasurg: Hierarchy-aware diffusion model for surgical video generation // International Conference on Medical Image Computing and Computer-Assisted Intervention. 2025. 310–319.

[4] *Blattmann Andreas, Dockhorn Tim, Kulal Sumith, Mendelevitch Daniel, Kilian Maciej, Lorenz Dominik, Levi Yam, English Zion, Voleti Vikram, Letts Adam, others* . Stable video diffusion: Scaling latent video diffusion models to large datasets // arXiv preprint arXiv:2311.15127. 2023.

[5] *Bruce Jake, Dennis Michael, Edwards Ashley, Parker-Holder Jack, Shi Yuge, Hughes Edward, Lai Matthew, Mavalankar Aditi, Steigerwald Richie, Apps Chris, others* . Genie: generative interactive environments // Proceedings of the 41st International Conference on Machine Learning. 2024. 4603–4623.

[6] *Cardoso M Jorge, Li Wenqi, Brown Richard, Ma Nic, Kerfoot Eric, Wang Yiheng, Murrey Benjamin, Myronenko Andriy, Zhao Can, Yang Dong, others* . Monai: An open-source framework for deep learning in healthcare // arXiv preprint arXiv:2211.02701. 2022.

[7] *Chen Boyuan, Martí Monsó Diego, Du Yilun, Simchowitz Max, Tedrake Russ, Sitzmann Vincent*. Diffusion forcing: Next-token prediction meets full-sequence diffusion // Advances in Neural Information Processing Systems. 2024. 37. 24081–24125.

[8] *Chen Chaofeng, Mo Jiadi*. IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. 2022.

[9] *Chen Ting, Kornblith Simon, Norouzi Mohammad, Hinton Geoffrey*. A simple framework for contrastive learning of visual representations // International conference on machine learning. 2020. 1597–1607.

[10] *Chen Tong, Yang Shuya, Wang Junyi, Bai Long, Ren Hongliang, Zhou Luping*. Surgsora: Object-aware diffusion model for controllable surgical video generation // International Conference on Medical Image Computing and Computer-Assisted Intervention. 2025. 521–531.

[11] *Chu Ruihang, He Yefei, Chen Zhekai, Zhang Shiwei, Xu Xiaogang, Xia Bin, Wang Dingdong, Yi Hongwei, Liu Xihui, Zhao Hengshuang, others* . Wan-move: Motion-controllable video generation via latent trajectory guidance // arXiv preprint arXiv:2512.08765. 2025.

[12] *Ciuti Gastone, Webster III Robert J, Kwok Ka-Wai, Menciassi Arianna*. Robotic surgery // Nature Reviews Bioengineering. 2025. 3, 7. 565–578.

[13] *Cui Justin, Wu Jie, Li Ming, Yang Tao, Li Xiaojie, Wang Rui, Bai Andrew, Ban Yuanhao, Hsieh Cho-Jui*. Self-Forcing++: Towards Minute-Scale High-Quality Video Generation // arXiv preprint arXiv:2510.02283. 2025.

[14] *Doersch Carl, Gupta Ankush, Markeeva Larisa, Recasens Adria, Smaira Lucas, Aytar Yusuf, Carreira Joao, Zisserman Andrew, Yang Yi*. Tap-vid: A benchmark for tracking any point in a video // Advances in Neural Information Processing Systems. 2022. 35. 13610–13626.

[15] *Ha David, Schmidhuber Jürgen*. World Models // arXiv preprint arXiv:1803.10122. 2018.

[16] *Hafner Danijar, Lillicrap Timothy, Ba Jimmy, Norouzi Mohammad*. Dream to control: Learning behaviors by latent imagination // arXiv preprint arXiv:1912.01603. 2019.

[17] *Hafner Danijar, Pasukonis Jurgis, Ba Jimmy, Lillicrap Timothy*. Mastering diverse domains through world models // arXiv preprint arXiv:2301.04104. 2023.

[18] *He Kaiming, Fan Haoqi, Wu Yuxin, Xie Saining, Girshick Ross*. Momentum contrast for unsupervised visual representation learning // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. 9729–9738.

[19] *He Xianglong, Peng Chunli, Liu Zexiang, Wang Boyang, Zhang Yifan, Cui Qi, Kang Fei, Jiang Biao, An Mengyin, Ren Yangyang, others* . Matrix-game 2.0: An open-source real-time and streaming interactive world model // arXiv preprint arXiv:2508.13009. 2025.

[20] *He Yufan, Guo Pengfei, Xu Mengya, Li Zhaoshuo, Myronenko Andriy, Imans Dillan, Liu Bingjie, Yang Dongren, Gu Mingxue, Ji Yongnan, others* . SurgWorld: Learning Surgical Robot Policies from Videos via World Modeling // arXiv preprint arXiv:2512.23162. 2025.

[21] *Henschel Roberto, Khachatryan Levon, Poghosyan Hayk, Hayrapetyan Daniil, Tadevosyan Vahram, Wang Zhangyang, Navasardyan Shant, Shi Humphrey*. Streamingt2v: Consistent, dynamic, and extendable long video generation from text // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. 2568–2577.

[22] *Huang Xun, Li Zhengqi, He Guande, Zhou Mingyuan, Shechtman Eli*. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion // arXiv preprint arXiv:2506.08009. 2025.

[23] *Huang Ziqi, He Yinan, Yu Jiashuo, Zhang Fan, Si Chenyang, Jiang Yuming, Zhang Yuanhan, Wu Tianxing, Jin Qingyang, Chanpaisit Nattapol, others* . Vbench: Comprehensive benchmark suite for video generative models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. 21807–21818.

[24] *Karaev Nikita, Makarov Yuri, Wang Jianyuan, Neverova Natalia, Vedaldi Andrea, Rupprecht Christian*. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025. 6013–6022.

[25] *Li Chenxin, Liu Hengyu, Liu Yifan, Feng Brandon Y, Li Wuyang, Liu Xinyu, Chen Zhen, Shao Jing, Yuan Yixuan*. Endora: Video generation models as endoscopy simulators // International conference on medical image computing and computer-assisted intervention. 2024. 230–240.

[26] *Li Wuyang, Pan Wentao, Luan Po-Chien, Gao Yang, Alahi Alexandre*. Stable Video Infinity: Infinite-Length Video Generation with Error Recycling // International Conference on Learning Representations. 2026.

[27] *Li Zhen, Zhu Zuo-Liang, Han Ling-Hao, Hou Qibin, Guo Chun-Le, Cheng Ming-Ming*. Amt: All-pairs multi-field transforms for efficient frame interpolation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. 9801–9810.

[28] *Liu Xinyu, Liu Hengyu, Wang Cheng, Liu Tianming, Yuan Yixuan*. Endogen: Conditional autoregressive endoscopic video generation // International Conference on Medical Image Computing and Computer-Assisted Intervention. 2025. 169–179.

[29] *Ning Mang, Li Mingxiao, Su Jianlin, Salah Albert Ali, Ertugrul Itir Onal*. Elucidating the Exposure Bias in Diffusion Models // 12th International Conference on Learning Representations, ICLR 2024. 2024.

[30] *Nwoye Chinedu Innocent, Elgohary Kareem, Srinivas Anvita, Zaid Fauzan, Lavanchy Joël L, Padoy Nicolas*. Cholectrack20: A multi-perspective tracking dataset for surgical tools // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025. 8942–8952.

[31] *Rapuri Sampath, Seenivasan Lalithkumar, Schneider Dominik, Soberanis-Mukul Roger, He Yufan, Ding Hao, Xu Jiru, Yu Chenhao, Jing Chenyan, Guo Pengfei, others* . SAW: Toward a Surgical Action World Model via Controllable and Scalable Video Generation // arXiv preprint arXiv:2603.13024. 2026.

[32] *Schmidgall Samuel, Kim Ji Woong, Kuntz Alan, Ghazi Ahmed Ezzat, Krieger Axel*. General-purpose foundation models for increased autonomy in robot-assisted surgery // Nature Machine Intelligence. 2024. 6, 11. 1275–1283.

[33] *Schmidt Florian*. Generalization in generation: A closer look at exposure bias // Proceedings of the 3rd Workshop on Neural Generation and Translation. 2019. 157–167.

[34] *Simonyan Karen, Zisserman Andrew*. Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. 2014.

[35] *Song Kiwhan, Chen Boyuan, Simchowitz Max, Du Yilun, Tedrake Russ, Sitzmann Vincent*. History-Guided Video Diffusion // International Conference on Machine Learning. 2025. 56242–56280.

[36] *Sun Wenqiang, Zhang Haiyu, Wang Haoyuan, Wu Junta, Wang Zehan, Wang Zhenwei, Wang Yunhong, Zhang Jun, Wang Tengfei, Guo Chunchao*. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling // arXiv preprint arXiv:2512.14614. 2025.

[37] *Turkcan Mehmet Kerem, Ballo Mattia, Filicori Filippo, Kostic Zoran*. Towards Suturing World Models: Learning Predictive Models for Robotic Surgical Tasks // arXiv preprint arXiv:2503.12531. 2025.

[38] *Wan Team, Wang Ang, Ai Baole, Wen Bin, Mao Chaojie, Xie Chen-Wei, Chen Di, Yu Feiwu, Zhao Haiming, Yang Jianxiao, others* . Wan: Open and advanced large-scale video generative models // arXiv preprint arXiv:2503.20314. 2025.

[39] *Wang Zile, Liu Zexiang, Li Jiaxing, Huang Kaichen, Xu Baixin, Kang Fei, An Mengyin, Wang Peiyu, Jiang Biao, Wei Yichen, others* . Matrix-Game 3.0: Real-Time and Streaming Interactive World Model with Long-Horizon Memory // arXiv preprint arXiv:2604.08995. 2026.

[40] *Wu Bing, Zou Chang, Li Changlin, Huang Duojun, Yang Fang, Tan Hao, Peng Jack, Wu Jianbing, Xiong Jiangfeng, Jiang Jie, others* . Hunyuanvideo 1.5 technical report // arXiv preprint arXiv:2511.18870. 2025.

[41] *Yang Zhuoyi, Teng Jiayan, Zheng Wendi, Ding Ming, Huang Shiyu, Xu Jiazheng, Yang Yuanming, Hong Wenyi, Zhang Xiaohan, Feng Guanyu, others* . CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer // arXiv preprint arXiv:2408.06072. 2024.

[42] *Zhang Lvmin, Cai Shengqu, Li Muyang, Wetzstein Gordon, Agrawala Maneesh*. Frame context packing and drift prevention in next-frame-prediction video diffusion models // arXiv preprint arXiv:2504.12626. 2025.

[43] *Zhang Richard, Isola Phillip, Efros Alexei A, Shechtman Eli, Wang Oliver*. The unreasonable effectiveness of deep features as a perceptual metric // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. 586–595.

[44] *Zia Aneeq, Berniker Max, Garcia Nespolo Rogerio, Perreault Conor, Bhattacharyya Kiran, Liu Xi, Wang Ziheng, Kondo Satoshi, Kasai Satoshi, Hirasawa Kousuke, others* . Intuitive Surgical SurgToolLoc Challenge Results: 2022-2023 // arXiv preprint arXiv:2305.07152v3. 2025.

Appendix

A Benchmark Evaluation Suite **14**

A.1 Instrument-Motion Accuracy and Tissue-Response Fidelity 14

A.2 Visual Quality 15

A.3 Temporal Consistency 16

B Extended Analysis **16**

B.1 Robustness and Metric Validation 16

B.2 Evaluation under Non-Contact Scenes and Optical Artifacts 17

B.3 Hyperparameter Sensitivity and Calibration 17

B.4 De Novo versus Autoregressive Generation 18

C Implementation Details **18**

D Limitations and Future Directions **19**

E Broader Impacts **20**

F Licenses for Existing Assets **20**

A Benchmark Evaluation Suite

Existing evaluation protocols primarily focus on pixel-level distributional similarity and perceptual fidelity, yet remain largely agnostic to the physical plausibility of generated dynamics. For surgical world modeling, faithful simulation of instrument motion and tissue response is equally vital. SurgWorld-Bench therefore establishes a rigorous evaluation framework along four complementary dimensions: **instrument-motion accuracy**, **tissue-response fidelity**, **visual quality**, and **temporal consistency**. Each dimension is formalized below alongside its metrics and computational protocols.

A.1 Instrument-Motion Accuracy and Tissue-Response Fidelity

Conventional holistic metrics such as FVD fail to decouple whether generated instruments adhere to prescribed controls or whether anatomical tissue exhibits plausible deformation upon contact. To quantitatively isolate these two orthogonal aspects, SurgWorld-Bench introduces **instrument-motion accuracy** and **tissue-response fidelity** as two trajectory-grounded dimensions evaluated over instrument and tissue regions, respectively. Both dimensions share an identical trajectory extraction and evaluation formulation, differing solely in the spatial domain over which trajectories are sampled.

Given a generated video and its ground-truth sequence, an off-the-shelf segmentation model[2] partitions the initial frame into instrument and tissue regions. A uniform 35×35 grid is sampled on this frame, with points assigned to the instrument set $\mathcal{P}_{\text{inst}}$ or tissue set $\mathcal{P}_{\text{tissue}}$ using the mask ($|\mathcal{P}_{\text{inst}}| = N_{\text{inst}}$, $|\mathcal{P}_{\text{tissue}}| = N_{\text{tissue}}$). CoTracker [24] propagates these initial points through both sequences, yielding trajectories $\{\mathbf{p}_t^{\text{pred}(i)}\}$ and $\{\mathbf{p}_t^{\text{gt}(i)}\}$ for each point i at frame $t \in [1, T]$. We use two metrics from TAP-Vid [14]: Average Jaccard (AJ) and Average Visible Position Accuracy ($\delta^{\text{vis}}_{\text{avg}}$).

Average Jaccard (AJ) jointly evaluates spatial localization accuracy and visibility prediction across distance thresholds $\Delta = \{1, 2, 4, 8, 16\}$ pixels:

$$\text{AJ} = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \frac{\sum_{i \in \mathcal{P}} \sum_{t=1}^{T} \text{TP}_{\delta,t}^{(i)}}{\sum_{i \in \mathcal{P}} \sum_{t=1}^{T} \left(\text{TP}_{\delta,t}^{(i)} + \text{FP}_{\delta,t}^{(i)} + \text{FN}_{\delta,t}^{(i)} \right)}, \tag{12}$$

where $\mathcal{P} \in \{\mathcal{P}_{\text{inst}}, \mathcal{P}_{\text{tissue}}\}$ denotes the evaluated point set, and T is the temporal horizon. Following TAP-Vid [14], a point is counted as $\text{TP}_{\delta,t}^{(i)}$ if correctly predicted as visible with $\|\mathbf{p}_t^{\text{pred}(i)} - \mathbf{p}_t^{\text{gt}(i)}\|_2 < \delta$, while position errors exceeding δ or visibility mismatches are counted as FP or FN. Counts are pooled across all evaluated trajectories and frames to produce a single score per clip.

Average Visible Position Accuracy ($\delta^{\text{vis}}_{\text{avg}}$) isolates spatial localization precision from occlusion classification, measuring the proportion of ground-truth visible points whose estimated positions fall within threshold δ:

$$\delta^{\text{vis}}_{\text{avg}} = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \frac{\sum_{i \in \mathcal{P}} \sum_{t=1}^{T} v_t^{(i)} \cdot \mathbb{1}\left[\|\mathbf{p}_t^{\text{pred}(i)} - \mathbf{p}_t^{\text{gt}(i)}\|_2 < \delta \right]}{\sum_{i \in \mathcal{P}} \sum_{t=1}^{T} v_t^{(i)}}, \tag{13}$$

where $v_t^{(i)} \in \{0, 1\}$ denotes the ground-truth visibility indicator for point i at frame t.

When computed over instrument points ($\mathcal{P}_{\text{inst}}$), AJ and $\delta^{\text{vis}}_{\text{avg}}$ quantify adherence to prescribed tool trajectories and motion. When evaluated over tissue points ($\mathcal{P}_{\text{tissue}}$), they measure reference-aligned deformation and retraction dynamics relative to the ground-truth video. This decoupled formulation enables fine-grained diagnosis, isolating control execution errors from implausible tissue responses.

To illustrate how the decoupled metrics distinguish instrument-motion accuracy from tissue-response fidelity, Figures 6 and 7 compare representative sequences under contrasting interaction regimes. With minimal instrument-tissue contact (Figure 6), instrument scores diverge sharply across cases, reflecting differences in trajectory quality, while tissue scores remain comparable due to similar background dynamics. During active manipulation (Figure 7), the metrics capture disparities in both tool trajectory tracking and tissue deformation fidelity. These complementary cases demonstrate how region-isolated evaluation decouples tool control execution from tissue response, providing diagnostic granularity beyond conventional holistic metrics.

[2] `https://github.com/Project-MONAI/model-zoo/tree/dev/models/endoscopic_tool_segmentation`

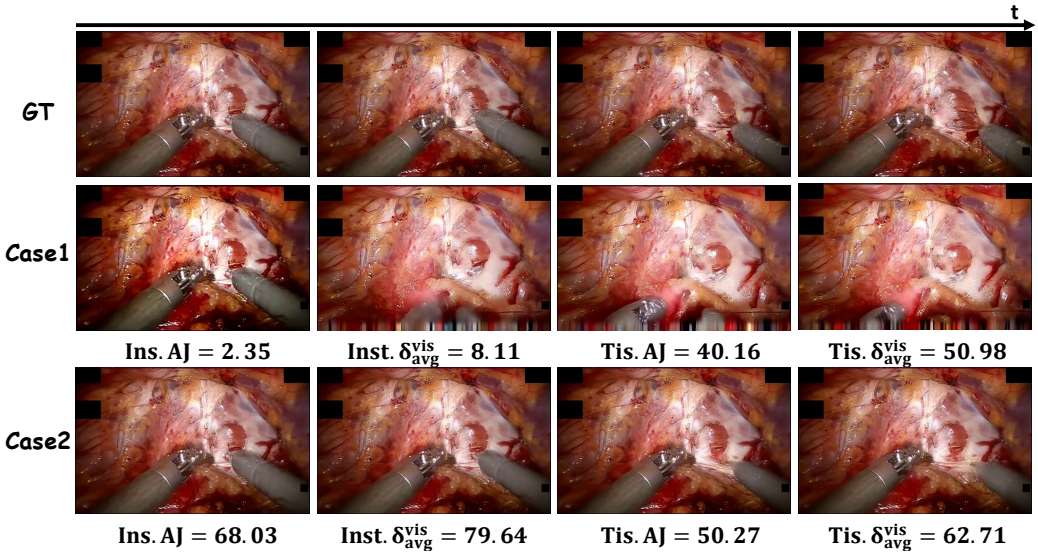

Figure 6: Decoupled metric behavior under minimal instrument-tissue contact. Instrument-region metrics diverge sharply to reflect substantial differences in tool trajectory tracking, while tissue-region scores remain comparable due to minimal deformation dynamics.

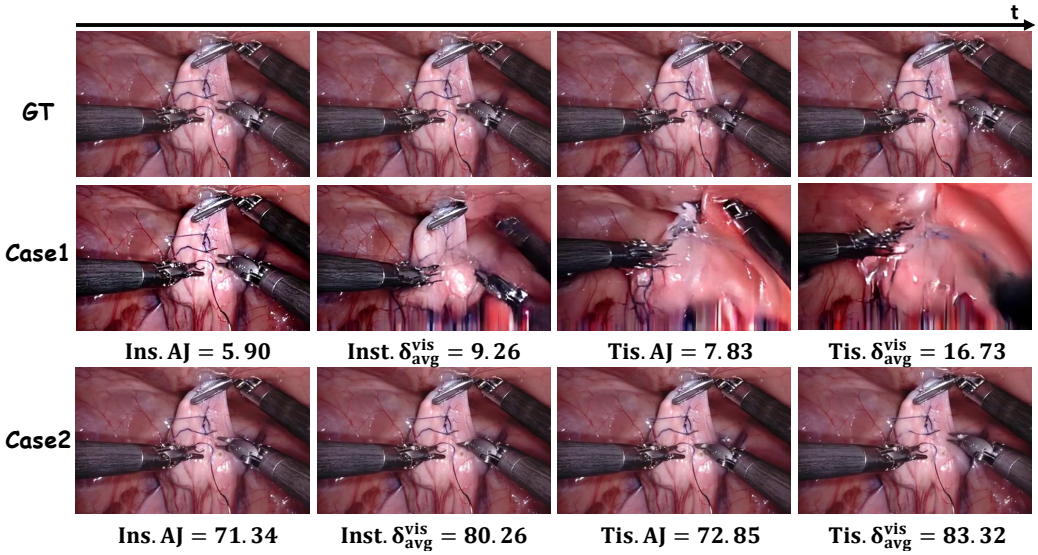

Figure 7: Decoupled metric behavior under active instrument-tissue interaction. Both instrument-motion and tissue-response scores capture the corresponding discrepancies in trajectory execution and contact deformation.

A.2 Visual Quality

Per-frame perceptual and structural fidelity is evaluated using three standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [43]. Specifically, PSNR measures low-level pixel reconstruction fidelity; SSIM evaluates structural preservation across luminance, contrast, and local spatial patterns; and LPIPS quantifies deep feature divergence within the latent space of an ImageNet-pretrained VGG backbone [34]. For a video of length T, each metric is computed frame-wise between the generated frame $\hat{f}_t$ and the ground truth f_t, and subsequently averaged over the sequence via the PyIQA library [8]. Higher PSNR and SSIM values indicate superior image reconstruction, whereas a lower LPIPS score reflects closer perceptual alignment.

A.3 Temporal Consistency

Following VBench [23], we evaluate temporal consistency using *Motion Smoothness*. A pretrained AMT frame interpolation model [27] reconstructs intervening frames from alternating generated frames. The score is computed as one minus the mean absolute pixel difference between reconstructed and corresponding generated frames, normalized by the pixel intensity range and averaged across evaluated frames. Higher scores indicate greater consistency with frame interpolation and smoother temporal transitions.

B Extended Analysis

B.1 Robustness and Metric Validation

Manual Annotation Alignment. To assess tracker fidelity, we manually annotate six corresponding trajectories per video clip (comprising three instrument and three tissue trajectories) across the reference, Wan2.1-Control, and SurgHorizon sequences, starting from identical initial points. Benchmarked against these manual ground-truth annotations, CoTracker [24] achieves an overall AJ of 86.57% and a visible position accuracy of 93.24%, yielding mean and median endpoint errors of 1.35 and 0.58 pixels, respectively, alongside a 95th-percentile error of 4.79 pixels. Re-evaluating the regional metrics on this manually annotated subset preserves SurgHorizon's consistent paired gains across all four key dimensions (Table 6). This confirms that tracker approximations do not introduce systematic bias, validating the fairness and robustness of our model comparisons.

Table 6: Trajectory evaluation benchmarked against manual annotations on the annotated subset. Scores are percentages; Δ denotes the paired gain of SurgHorizon over Wan2.1-Control.

Trajectory Source	Method	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$
CoTracker [24]	Wan2.1-Control	58.12	69.95	49.77	59.58
	SurgHorizon	64.40	75.78	54.25	64.40
	Paired Δ	+6.28	+5.82	+4.48	+4.81
Manual Annotation	Wan2.1-Control	60.84	69.45	49.26	58.32
	SurgHorizon	66.88	75.02	53.76	63.02
	Paired Δ	+6.04	+5.57	+4.50	+4.70

Perturbation Stress-Testing. We test robustness against synthetic tracking errors and mask misalignments. Across 1,000 Monte Carlo runs with coordinate jitter and visibility flips, SurgHorizon consistently retains positive paired advantages under moderate and severe tracking perturbations (Table 7). Similarly, morphologically eroding and dilating the initial tool masks yields strictly positive lower bounds across all bootstrap confidence intervals. These stress tests confirm that benchmark conclusions and rankings remain robust under tracking noise and boundary segmentation variations.

Table 7: Sensitivity of paired gains (Δ, percentage points) under tracking and mask perturbations. Brackets denote 95% Monte Carlo percentiles (noise rows) or paired clip-bootstrap CIs (mask rows).

Perturbation Regime	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$
Clean (Default Protocol)	+6.42	+6.35	+5.50	+5.78
Tracking: Moderate Noise	+5.83 [5.78, 5.87]	+6.14 [6.10, 6.17]	+5.10 [5.08, 5.12]	+5.61 [5.59, 5.62]
Tracking: Severe Noise	+5.71 [5.64, 5.78]	+5.87 [5.80, 5.93]	+5.34 [5.31, 5.37]	+5.54 [5.51, 5.56]
Mask: Erosion (2 px)	+6.50 [5.88, 7.15]	+6.38 [5.84, 6.93]	+5.49 [4.40, 6.66]	+5.78 [4.69, 6.95]
Mask: Erosion (8 px)	+6.60 [5.87, 7.35]	+6.48 [5.88, 7.07]	+5.49 [4.51, 6.58]	+5.78 [4.77, 6.87]
Mask: Dilation (2 px)	+6.33 [5.79, 6.91]	+6.36 [5.86, 6.89]	+5.51 [4.43, 6.69]	+5.78 [4.69, 6.96]
Mask: Dilation (8 px)	+5.99 [5.50, 6.51]	+6.22 [5.73, 6.76]	+5.53 [4.42, 6.77]	+5.79 [4.67, 6.97]

B.2 Evaluation under Non-Contact Scenes and Optical Artifacts

Non-Contact Scenes. In surgical procedures, tissue displacement often arises from camera ego-motion and physiological shifts rather than direct instrument manipulation. To evaluate dynamics prediction when external tool forcing is absent, we select a subset of clips exhibiting visible instruments without observable contact. In this regime, tissue trajectories capture passive anatomical and camera-induced motion relative to the reference sequence. As reported in Table 8, SurgHorizon outperforms Wan2.1-Control across all four regional dimensions, verifying that our model maintains accurate spatial modeling even without contact-driven deformation priors.

Natural Imaging Artifacts. Endoscopic video feeds frequently suffer from transient visual degradation, including electrocautery smoke, motion blur, and lens contamination. We evaluate a challenging short-horizon subset exhibiting these natural artifacts, where anatomical structures remain trackable. Despite degraded visual evidence, SurgHorizon consistently retains its competitive margin over Wan2.1-Control across all instrument and tissue metrics (Table 8). These evaluations confirm that SurgHorizon's motion fidelity generalizes beyond idealized, artifact-free surgical conditions.

Table 8: Regional evaluation under non-contact scenes and optical imaging artifacts. Scores are percentages; Δ denotes the paired gain of SurgHorizon over Wan2.1-Control.

Evaluation Regime	Method	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$
No Observable Contact	Wan2.1-Control	39.23	53.16	41.17	52.69
	SurgHorizon	44.36	57.21	45.95	57.29
	Paired Δ	+5.13	+4.05	+4.78	+4.60
Natural Artifacts	Wan2.1-Control	35.52	48.61	32.61	45.09
	SurgHorizon	39.84	53.91	38.08	51.05
	Paired Δ	+4.32	+5.30	+5.47	+5.96

B.3 Hyperparameter Sensitivity and Calibration

DCR Loss Weight. We vary λ_{DCR} on the short-horizon set (Table 9). Performance remains stable, with $\lambda_{\text{DCR}} = 0.01$ achieving the best scores on all four metrics. Lower weights weaken dynamic regularization, while larger values over-constrain feature optimization without improving trajectory adherence.

Table 9: Sensitivity to the DCR loss weight λ_{DCR} on the short-horizon evaluation set.

λ_{DCR}	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$
0.005	44.95	57.59	41.08	54.66
0.01	45.76	58.43	41.68	55.29
0.02	45.35	58.04	40.61	53.94

DAT Mode Probabilities. Drift Adaptation Training balances probe exposures, photometric perturbations, and clean conditioning via mixture probabilities $(p_{\text{probe}}, p_{\text{aug}}, p_{\text{clean}})$. As shown in Table 10, reducing probe exposure to 0.2 leads to clear accuracy drops across both instruments and tissue, confirming the importance of autoregressive error feedback during training. The balanced configuration $(0.5, 0.4, 0.1)$ maintains strong tracking fidelity while mitigating exposure drift.

Table 10: Sensitivity to DAT mode sampling probabilities on the long-horizon evaluation set.

$(p_{\text{probe}}, p_{\text{aug}}, p_{\text{clean}})$	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$
$(0.7, 0.2, 0.1)$	27.60	39.95	24.39	38.03
$(0.5, 0.4, 0.1)$	27.13	39.61	24.15	38.19
$(0.2, 0.7, 0.1)$	26.63	39.25	23.90	37.84

Photometric Calibration. To counter long-horizon visual degradation, the photometric perturbation ranges are empirically calibrated to cover distribution drift observed during autoregressive rollouts. Table 11 evaluates model robustness across varying augmentation magnitudes. The calibrated default ranges achieve the best balance across both instrument and tissue metrics; weaker perturbations provide insufficient regularization against accumulated drift, whereas overly aggressive transformations introduce severe visual distortions that degrade trajectory adherence.

Table 11: Photometric augmentation sensitivity on the long-horizon evaluation set. Intervals define augmentation sampling bounds; noise indicates Gaussian standard deviation.

Setting	Brightness	Contrast	Sharpness	Noise	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$
Weak	[0.95,1.15]	[1.0,1.5]	[0.85,1.25]	[0,0.01]	26.85	39.65	24.13	37.88
Default	[0.8,1.5]	[1.0,3.0]	[0.5,2.0]	[0,0.05]	27.13	39.61	24.15	38.19
Strong	[0.6,2.0]	[1.0,5.0]	[0.2,3.0]	[0,0.10]	26.36	38.51	24.01	38.07

B.4 De Novo versus Autoregressive Generation

We evaluate autoregressive rollout against de novo synthesis on long-horizon clips (Table 12). For both methods, restarting generation between segments mitigates error propagation, thereby improving instrument motion accuracy and perceptual quality (LPIPS). However, for SurgHorizon, each restart abruptly discards the evolving interaction history, impairing tissue-response fidelity and temporal motion smoothness. These findings highlight a fundamental trade-off between suppressing cumulative drift and preserving dynamic interaction states. Correcting visual drift without resetting these states or relying on future reference frames remains an open direction for future work.

Table 12: Comparison between de novo and autoregressive generation on long-horizon clips.

Method	Generation	Inst. AJ	Inst. $\delta^{\text{vis}}_{\text{avg}}$	Tissue AJ	Tissue $\delta^{\text{vis}}_{\text{avg}}$	LPIPS	Smoothness
Wan2.1-Control	Autoregressive	18.92	28.30	19.91	31.90	0.4219	0.99510
	De novo	22.44	36.15	20.81	34.64	0.3675	0.99248
SurgHorizon	Autoregressive	27.13	39.61	24.15	38.19	0.3712	0.99530
	De novo	31.68	46.75	21.85	36.20	0.3494	0.99311

Additional Visualizations. More visualization results are provided in the supplementary material.

C Implementation Details

Data Curation Pipeline. Training data are sourced from three publicly available surgical video datasets: GraSP [2], CholecTrack20 [30], and SurgToolLoc22 [44], and processed through a unified curation pipeline as follows. Raw surgical videos are first resampled to 24 frames per second (fps) and segmented into non-overlapping clips of 81 frames each. Each clip is then center-cropped and resized to a spatial resolution of 512×320 to match the input specification of the base generative model. All clips undergo manual inspection to remove severe blur, smoke, or lens contamination that prevents reliable trajectory extraction, content irrelevant to the surgical procedure (e.g., external camera views or idle periods), and predominantly static scenes with negligible instrument or tissue motion. Moderate imaging artifacts are retained when the relevant motion remains observable. After filtering, the final training set comprises 10,687 clips in total: 3,104 from GraSP, 3,360 from CholecTrack20, and 4,223 from SurgToolLoc22.

SurgHorizon. SurgHorizon is built upon Wan2.1-1.3B[3], an image-to-video control variant of Wan2.1 [38] that extends the base flow-matching video diffusion transformer with image conditioning and control capabilities. The backbone is pretrained on large-scale open-domain video data. SurgHorizon is trained on the curated surgical training set with LoRA fine-tuning. Training is conducted on eight NVIDIA RTX A6000 48GB GPUs for a total of 100K iterations. Mixed-precision training with

[3] `https://huggingface.co/alibaba-pai/Wan2.1-Fun-V1.1-1.3B-Control`

bf16 is adopted to balance memory efficiency and training stability. The full training configuration and module-specific hyperparameters are summarized in Table 13.

Table 13: Implementation details of SurgHorizon.

Hyperparameter	Value	Description
Training Configuration		
Optimizer	AdamW	Optimizer in the training process
Learning rate	1×10^{-4}	Learning rate during training
Weight decay	1×10^{-2}	L2 regularization coefficient
LR scheduler	Constant	Learning rate schedule policy
Total training steps	100K	Optimization iterations
Precision	bf16	Training precision
Gradient accumulation	1	Micro-batches per update
Gradient checkpointing	false	Memory-compute tradeoff
LoRA rank	32	Low-rank adaptation rank
LoRA target modules	q, k, v, o, ffn.0, ffn.2	Layers adapted with LoRA
Frame Width	512	Width of training frames
Frame Height	320	Height of training frames
Number of frames	81	Frames per training clip
Drift Adaptation Training (DAT)		
Model-intrinsic proxy prob.	0.5	Probability of self-predicted error proxy
Model-agnostic proxy prob.	0.4	Probability of photometric error proxy
Clean mode prob.	0.1	Probability of unperturbed training
Brightness range	[0.8, 1.5]	Brightness jitter range
Contrast range	[1.0, 3.0]	Contrast jitter range
Sharpness range	[0.5, 2.0]	Sharpness jitter range
Gaussian noise σ	[0, 0.05]	Additive noise standard deviation range
Deformation Consistency Regularization (DCR)		
Loss weight λ_{DCR}	0.01	DCR loss weight
Temperature γ	0.1	Temperature for contrastive loss
Tracked points	64	Number of trajectories for supervision
Projector hidden dim	512	MLP projector intermediate dimension
Projector output dim	256	MLP projector output dimension
Feature source	all	Regularized regions

D Limitations and Future Directions

While SurgHorizon achieves compelling fidelity across complex surgical scenarios, several modeling and representational limitations remain.

2D Visual Trajectories vs. True Robot Kinematics. Our current control conditioning relies on 2D visual point trajectories extracted from videos rather than true robot kinematics, such as 6-DoF end-effector poses or joint configurations. While this formulation enables training on diverse *in-vivo* surgical datasets lacking telemetry data, 2D trajectories inherently suffer from projective ambiguities and lack 3D spatial awareness. Conditioning world models on calibrated 3D end-effector transformations or robotic joint states represents a crucial next step toward deployment in closed-loop policy execution.

Unobserved Interaction Forces. The current framework does not explicitly model physical interaction cues, such as applied forces, contact torques, or tactile feedback. Consequently, interactions that share identical initial visuals and 2D tool trajectories but involve distinct applied forces remain indistinguishable to the model. Incorporating multi-modal force and tactile measurements will be vital to resolving such dynamic ambiguities.

Extreme Non-Rigid Deformations and Delicate Sutures. Faithfully simulating large-scale non-rigid tissue deformations remains non-trivial (Fig. 8, 1). Severe anatomical manipulation often exposes occluded cavities and novel textures absent from the context frames, requiring generative hallucination beyond the observable visual evidence. In addition, fine-grained structures such as suture threads frequently exhibit temporal flickering or blurriness (Fig. 8, 2), largely due to the spatial compression bottlenecks of video VAE. Exploring multi-scale architectures or domain-specific autoencoders presents a promising direction to mitigate this fine-grained information loss.

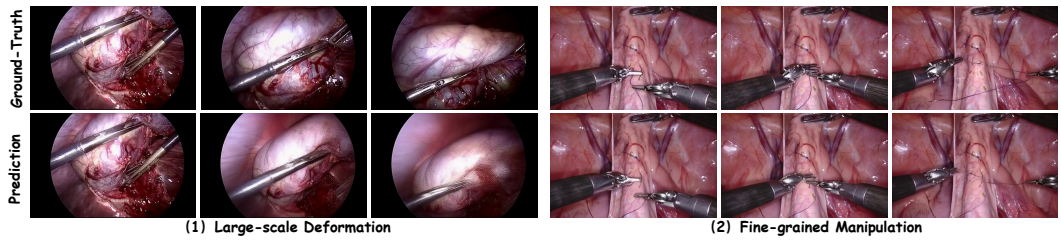

Figure 8: Representative failure cases of SurgHorizon. (1) Extreme tissue traction reveals unobserved internal surfaces beyond the temporal conditioning context. (2) Delicate suture thread dynamics degrade due to spatial latent compression bottlenecks.

E Broader Impacts

This work introduces a surgical world model capable of simulating long-horizon operative dynamics, presenting notable benefits for clinical robotics alongside essential safety considerations. On the one hand, synthesizing diverse surgical scenarios on demand lowers reliance on scarce, privacy-sensitive intraoperative recordings, facilitating scalable surgical education, preoperative planning, and scalable synthetic data generation for downstream robotic perception and policy learning. On the other hand, generative models that produce visually convincing yet biomechanically unverified dynamics pose inherent clinical risks if deployed naively. Hallucinated anatomical structures or unfaithful tissue responses could reinforce flawed surgical intuition in trainees or induce erratic robotic behavior without rigorous sim-to-real transfer validation. To mitigate these concerns, simulated surgical rollouts must be explicitly designated as synthetic, verified by expert clinicians before safety-critical use, and distributed through gated research frameworks under strict human-in-the-loop oversight.

F Licenses for Existing Assets

This work builds upon several publicly available datasets and pretrained models. GraSP [2] is released under the MIT License by BCV-Uniandes. SurgToolLoc22 [44] is publicly released by Intuitive Surgical as part of the MICCAI EndoVis 2022 Challenge for research use. CholecTrack20 [30] is released under a research-only Data Use Agreement (DUA) by the CAMMA research group, University of Strasbourg, under CC-BY-NC-SA 4.0. The Wan2.1 foundation model [38] is released under the Apache 2.0 License. CoTracker [24] is released under the CC-BY-NC 4.0 License by Meta Research. All assets are used in accordance with their respective terms of use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions are clearly outlined in both the abstract and the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical results, theorems, or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Comprehensive implementation details are provided in the experiment section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and pretrained models will be publicly released. The SurgWorld-Bench is constructed from publicly available datasets, and detailed data preparation instructions are provided in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and evaluation details are specified in the experiment section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported due to the prohibitive computational cost of repeated training runs for world models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: GPU type, memory, and training time are reported in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in the appendix, covering both the potential of long-horizon surgical video generation for training and simulation, and risks such as generating unrealistic surgical content that could mislead clinical decision-making.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The model will be released under a research-only license with usage guidelines that restrict application to non-clinical settings.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used in SurgWorld-Bench (GraSP, SurgToolLoc22, Cholec-Track20) and the base model (Wan2.1) are properly cited. The respective licenses are acknowledged in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The proposed SurgWorld-Bench and the pretrained model are documented with construction details, evaluation protocols, and dataset statistics. Both will be released with accompanying documentation and usage instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects. All surgical video data used are sourced from existing publicly available datasets with their own ethical approvals.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology of this work does not involve large language models as a component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.