MEDICAL WORLD MODEL: GENERATIVE SIMULATION OF LONG-HORIZON SURGERY FROM UNTIMED ACTION DESCRIPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Surgical descriptions convey a surgeon's intended actions during procedures that can span hours, yet they leave open how those actions unfold and reshape the operative scene over time. Translating such descriptions into sustained visual simulation requires capturing both the progression of interacting actions and their lasting effects on tissue. We introduce SurgWAVE, a world model that generates successive surgical video blocks from an initial observation and untimed action descriptions. The model reconciles correlated descriptions without prescribing a frame-wise action schedule, retains accumulated intervention effects through utility-calibrated recurrent memory, and preserves local structure through training-only spatial distillation. To support this task, we introduce LiRAct, an expert-annotated laparoscopic liver-resection dataset and benchmark comprising 3,333 operative episodes from 18 patients. Experiments on LiRAct and Cholec80 demonstrate state-of-the-art performance in conditional surgical video generation. The code and benchmark will be released upon acceptance.

1 INTRODUCTION

Medical world models capture how clinical interventions shape future patient states through generative simulation. For example, MeWM builds treatment-conditioned tumor evolution (Yang et al., 2025). Extending this action-conditioned perspective to the operating room, a *surgical world model* simulates how an operative scene unfolds under described surgical actions, capturing instrument motion and tissue changes such as deformation and dissection.

Unlike radiotherapy, a surgery usually takes hours, and thus requires *long-horizon modeling*. Under such circumstances, actions have persistent, interdependent effects: retraction exposes tissue for dissection, while dissection changes the anatomy encountered by subsequent actions. Simulating these processes requires preserving accumulated tissue changes beyond the model's attention window while maintaining instrument shape and tool–tissue contact across successive predictions.

Recent methods condition surgical video generation on action text, tool trajectories, robot kinematics, or spatial controls (He et al., 2025; Rapuri et al., 2026; Li et al., 2026; Zhou et al., 2026). However, surgical descriptions may specify what happens without prescribing exactly when. For example, *the grasper retracts tissue while the dissector separates it* (Figure 1(a,b)) describes overlapping events but leaves their onset and duration unspecified; retraction may pause and resume as dissection proceeds. We call these *untimed action descriptions*. They allow users to describe a procedure in the absence of frame-level action intervals or instrument trajectories. Although language can manipulate surgical video generation (He et al., 2025), generating long sequences from such inputs requires resolving how each action influences the current prediction as the overlaps and the scene evolves. We address this challenge through long-horizon surgical simulation from untimed action descriptions.

To study this setting, we introduce **LiRAct**, an expert-annotated dataset and benchmark for laparoscopic **Li**ver **R**esection **Act**ion-conditioned video generation. It comprises 3,333 operative episodes from 18 patients and 5,692 camera-view records across paired-view and monocular episodes. Annotations describe instruments and events, with tissue targets and image locations where available.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

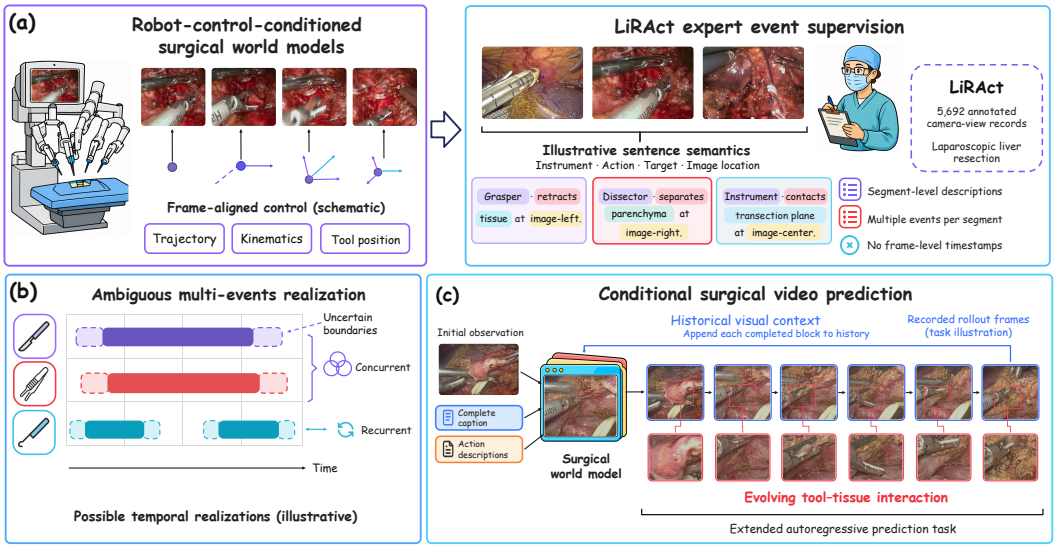

Figure 1: Surgical video generation from untimed action descriptions. (a) Frame-aligned robot controls and the expert descriptions used in LiRAct. Descriptions identify instruments and events, with target and image-location information where available. (b) The same descriptions can correspond to different event intervals, overlaps, and repetitions. (c) The model predicts successive video blocks from an initial observation, a complete caption, and action descriptions extracted from the caption. Completed blocks become visual history for subsequent predictions.

We also present **SurgWAVE**, a **Surg**ical **W**orld model for **A**ction-conditioned **V**ideo **E**volution. From an initial observation, a complete caption, and its extracted action descriptions, SurgWAVE generates successive video blocks using three mechanisms:

- **Correlation-Aware Signed Action Conditioning** combines state-dependent flow corrections from correlated descriptions through closed-form determinantal aggregation. Signed values permit additive and subtractive flow corrections without fixed action timing.
- **Contractive Residual State Memory (CReSM)** summarizes changes relative to the initial scene in a bounded recurrent state, independently of the rolling cache.
- **Predicted-State Spatial Distillation** uses a frozen LingBot-Vision observer (Fu et al., 2026) to constrain local relations at the predicted clean state through a bounded auxiliary gradient. The observer is used only during training.

SurgWAVE achieves the lowest reported FVD among the evaluated methods on LiRAct and Cholec80 under their documented settings. Component analyses and expert assessment further examine generation quality and activity compatibility.

2 Related Work

Surgical world models. Cosmos-H-Surgical, VISAGE, and SG2VID condition surgical video generation on action text, action triplets, or scene graphs (He et al., 2025; Yeganeh et al., 2024; Sivakumar et al., 2026). Other methods provide spatial or motion guidance through affordance masks and tool-tip trajectories, articulated kinematics, hierarchical masks, or instrument keypoints (Rapuri et al., 2026; Li et al., 2026; Zhou et al., 2026; Pan et al., 2026). SurgVeo emphasizes expert assessment, whereas SurgWMBench studies instrument-motion planning (Chen et al., 2025; Liu et al., 2026). SurgWAVE addresses generation from untimed descriptions; LiRAct evaluates their compatibility with generated liver-resection videos.

Autoregressive generation and memory. Self-forcing reduces the mismatch between training histories and autoregressive inference (Huang et al., 2025). Rolling Forcing, LongLive, and

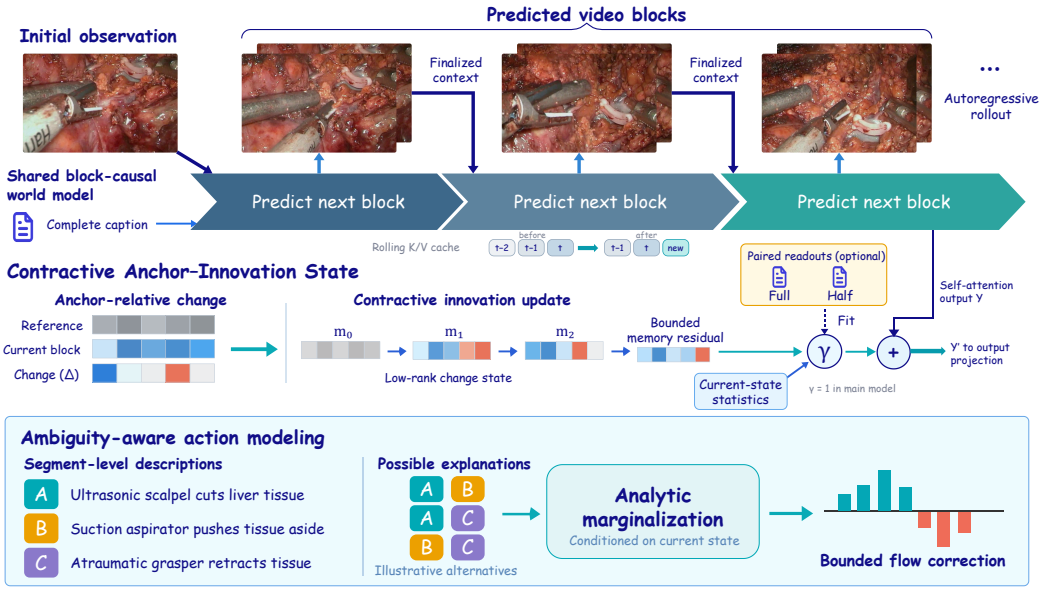

Figure 2: **Overview of SurgWAVE.** Successive video blocks are generated from the complete caption, extracted action descriptions, and visual history. The action module combines correlated descriptions into a bounded flow correction. CReSM summarizes changes relative to the initial scene and conditions self-attention through a bounded residual. Action subsets are schematic; recorded frames illustrate the autoregressive rollout.

LongLive-2.0 develop efficient streaming generation (Liu et al., 2025; Yang et al., 2026; Chen et al., 2026b). Episode-specific representations, state-space models, memory tokens, and retrieval provide complementary ways to retain history (Hong et al., 2025; Yu et al., 2025; Kim et al., 2026; Hu et al., 2026; Zhao et al., 2026). CReSM maintains a bounded recurrent summary of changes relative to the initial scene, independently of the rolling cache, with one committed update per video block.

Structural supervision. CTRL-Adapter, MVideo, and Through-The-Mask use explicit structural conditions, while UnityVideo jointly learns generation and auxiliary tasks (Lin et al., 2025; Zhou et al., 2024; Yariv et al., 2025; Huang et al., 2026). Endora, SurgSora, and SAM2VideoX exploit pretrained features or their relations (Li et al., 2024; Chen et al., 2026a; Fei et al., 2025). Our predicted-state spatial distillation uses the Jacobian of a frozen LingBot-Vision observer (Fu et al., 2026) to induce a local metric on the final action-corrected flow error, yielding a bounded auxiliary gradient without teacher computation at inference.

3 Method

SurgWAVE is a block-autoregressive flow-matching model conditioned on an initial observation and untimed action descriptions. It combines correlation-aware signed action conditioning, Contractive Residual State Memory (CReSM), and predicted-state spatial distillation (Figure 2).

3.1 Problem Definition

Let $x_0^{1:B}$ denote a video latent partitioned into B blocks. Inputs comprise a complete caption c, an indexed multiset of action descriptions $\mathcal{A} = [a_i]_{i=1}^{N}$, and history H_b containing the observed prefix and generated blocks. Descriptions have no frame-aligned onset or duration. Repetitions retain separate indices; the backbone independently encodes the complete caption.

The generator uses Wan2.2-TI2V-5B (Team Wan et al., 2025) and a block-causal architecture (Chen et al., 2026b). For clean latent x_0, noise $\epsilon \sim \mathcal{N}(0, I)$, and noise level $\sigma \in [0, 1]$, define $x_\sigma = (1 - \sigma)x_0 + \sigma\epsilon$ and target $v^\star = \epsilon - x_0$. The memory-conditioned backbone predicts v_θ^{native};

action conditioning adds Δv_ϕ, giving $\widehat{v} = v_\theta^{\text{native}} + \Delta v_\phi$. Here θ comprises trainable backbone and memory parameters, and ϕ the action projections. With $e = \widehat{v} - v^\star$, training minimizes

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\big[\|e\|_W^2\big], \tag{1}$$

where the expectation is over training clips, noise levels, and Gaussian noise. Matrix expressions use vectorized tensors; $\|e\|_W^2 = e^\top W e$, with diagonal $W \succeq 0$ incorporating timestep weights, valid-frame masks, and normalization. The clean prediction is $\widehat{x}_0 = x_\sigma - S_\sigma \widehat{v}$, where diagonal S_σ contains coordinate-wise noise scales, allowing different levels across blocks.

3.2 Correlation-Aware Signed Action Conditioning

State-dependent representations. Trainable projections map frozen, pooled text embeddings to keys $k_i \in \mathbb{R}^d$ and signed values $r_i \in \mathbb{R}^C$, with key width d and latent channel dimension C. An affine query map (W_q, b_q) gives $q_{t,p} \in \mathbb{R}^d$ at latent frame t and VAE-grid position p:

$$q_{t,p} = W_q[x_{\sigma,t,p}; v_{\theta,t,p}^{\text{native}}; \sigma_t] + b_q, \tag{2}$$

$$w_{t,i} = 0.25 + \text{sigmoid}\Big(\bar{q}_t^\top k_i/\sqrt{d}\Big), \qquad \bar{q}_t = \frac{1}{P}\sum_{p=1}^{P} q_{t,p}. \tag{3}$$

Pooling over P positions gives the relevance weights $w_{t,i}$. Set $K_t = [\sqrt{w_{t,i}}k_i]_{i=1}^N \in \mathbb{R}^{d\times N}$ and $R_t = [\sqrt{w_{t,i}}r_i]_{i=1}^N \in \mathbb{R}^{C\times N}$.

Determinantal aggregation. Let $G_t = K_t^\top K_t$. For subset $S \subseteq \{1,\dots,N\}$, $K_{t,S}$ and $R_{t,S}$ select its columns, and $G_{t,SS}$ is the corresponding principal submatrix. Define the subset probability and response by

$$\pi_t(S) = \frac{\det(G_{t,SS}/\lambda)}{\det(I_N + G_t/\lambda)}, \qquad d_{S,t,p} = R_{t,S}(G_{t,SS})^\dagger K_{t,S}^\top q_{t,p}, \tag{4}$$

where I_N is the identity, $\lambda > 0$ controls regularization and subset size, and $\dagger$ is the Moore–Penrose pseudoinverse. The empty subset has determinant one and response zero. Gram determinants measure squared key volumes: dependent subsets have zero weight, favoring complementary representations without imposing mutual exclusion between clinical actions. Determinantal sampling (Fanuel et al., 2021) yields

$$\bar{d}_{t,p} = \mathbb{E}_{S\sim\pi_t}[d_{S,t,p}] = R_t(G_t + \lambda I_N)^{-1} K_t^\top q_{t,p}. \tag{5}$$

One regularized solve per frame serves all spatial queries. Appendix G proves the identity for signed values and rank-deficient keys. Related to Intention (Garnelo & Czarnecki, 2023), this operator directly conditions flow. Recomputed queries allow overlapping and recurring contributions without estimating event boundaries; the autoregressive generator learns event order.

Bounded correction. A common scale bounds $\Delta v_{\phi,t,p} = \alpha_t \bar{d}_{t,p}$ relative to the native prediction:

$$\widehat{v}_{t,p} = v_{\theta,t,p}^{\text{native}} + \alpha_t \bar{d}_{t,p}, \qquad \alpha_t = \min\left(1, \frac{\eta\ \text{sg}[\text{RMS}(v_{\theta,t}^{\text{native}})]}{\sqrt{\text{mean}(\bar{d}_t^2) + \varepsilon}}\right), \tag{6}$$

where sg stops gradients, $\eta > 0$ sets the budget, and $\varepsilon > 0$ stabilizes division. RMS and mean cover frame channels and spatial positions. Correction applies only to unobserved entries at positive noise, identically in training and sampling. For fixed backbone features and queries, aggregation is permutation-invariant and preserves multiplicity.

3.3 Contractive Residual State Memory (CReSM)

Anchored dynamics. CReSM retains scene changes beyond the rolling attention window using a first-frame anchor and recurrent state. For head width d_h and state dimension d_m, value features $V_{t,p} \in \mathbb{R}^{d_h}$ are pooled over attention-token positions p and projected by $E \in \mathbb{R}^{d_m \times d_h}$ as $p_t = E\ \text{mean}_p(V_{t,p})$. Omitting layer and head indices, the anchor is $a = p_{\text{first}}^{\text{clean}}$, and the innovation and candidate state are

$$u_b = \frac{1}{|\mathcal{B}_b|}\sum_{t\in\mathcal{B}_b} \tanh(p_t - a), \qquad \widetilde{m}_b = \rho m_{b-1} + (1-\rho)u_b, \tag{7}$$

where $\mathcal{B}_b$ indexes block b's latent frames, $a, u_b, \widetilde{m}_b, m_b \in \mathbb{R}^{d_m}$, $m_0 = 0$, and learned retention satisfies $0 < \rho < \rho_{\max} < 1$. Bounded innovations ensure $\|m_b\|_\infty \le 1$; for a common innovation sequence, the recurrence contracts with factor ρ (Appendix G).

Bounded readout. A decoder $U \in \mathbb{R}^{d_h \times d_m}$ produces a spatially broadcast residual before the attention output projection:

$$\begin{aligned} \mathcal{R}_t &= \epsilon_m s_t \tanh(U\widetilde{m}_b), \qquad Y'_t = Y_t + \mathcal{R}_t, \\ s_t &= \mathrm{sg}[\max(\mathrm{RMS}(Y_t), \varepsilon_{\mathrm{rms}})]\,, \end{aligned} \tag{8}$$

Here Y_t and Y'_t are the native and memory-conditioned attention outputs, $\epsilon_m > 0$ controls readout strength, and $\varepsilon_{\mathrm{rms}} > 0$ is a numerical floor. RMS covers spatial positions, heads, and channels; the residual satisfies $\|\mathcal{R}_t\|_\infty \le \epsilon_m s_t$.

Block-causal propagation. Each denoising evaluation forms a candidate from the same committed state m_{b-1}. After finalization, clean features are recomputed and m_b is committed once. The first clean commit fixes a; earlier evaluations use the current first-frame feature provisionally. Anchor and state survive cache eviction. During teacher forcing, noisy blocks branch from preceding clean states, whose recurrence remains differentiable within the sequence; noisy candidates never enter subsequent history. At inference, commits are detached between blocks. Memory thus advances in video time, independently of denoising evaluations.

3.4 Predicted-State Spatial Distillation

A frozen LingBot-Vision ViT-small teacher (Fu et al., 2026), distilled from a larger model pretrained with masked boundary modeling, defines local structural relations. The differentiable observer $\mathcal{O}$ combines the frozen VAE decoder, teacher, and relation extraction. It decodes a latent suffix and returns $1 - \cos(z_i, z_j)$ for neighboring normalized patch features z_i, z_j from its final RGB frame, excluding the outer token border. Indices i, j denote horizontally or vertically adjacent patches.

Local spatial metric. Restrict quantities to an eligible low-noise suffix with n_x scalar latent coordinates and common weight $W_{\mathrm{suffix}} = \omega I_{n_x}, \omega \ge 0$. For n_r retained relations, $\mathcal{O} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_r}$. At detached reference $\bar{x}_0 = \mathrm{sg}[x_\sigma - S_\sigma \widehat{v}]$, define

$$J = D\mathcal{O}(\bar{x}_0)(-S_\sigma), \qquad h = \frac{J^\top J e}{\tau}, \qquad \tau = \max\left(\frac{1}{n_\xi}\sum_{k=1}^{n_\xi} \|J^\top \xi_k\|_2^2, \varepsilon_\tau\right), \tag{9}$$

Here $D\mathcal{O}$ is the observer's latent-input Jacobian, $J \in \mathbb{R}^{n_r \times n_x}$, $e, h \in \mathbb{R}^{n_x}$, and $\varepsilon_\tau > 0$ is a numerical floor. Each of the $n_\xi \ge 1$ Rademacher probes $\xi_k \in \{-1, +1\}^{n_r}$ has independent equiprobable signs. They estimate $\mathrm{tr}(J^\top J)$ (Hutchinson, 1989); Jacobian products evaluate $J^\top J e$ without constructing J. Detachment fixes the reference while retaining observer input derivatives.

With J, τ, ω fixed, the local quadratic $\lambda_{\mathrm{pb}}\omega\|Je\|_2^2/\tau$ has gradient $2\lambda_{\mathrm{pb}}\omega h$, supplied at the reference prediction by the stopped linear surrogate

$$\mathcal{L}_{\mathrm{pb}} = 2\lambda_{\mathrm{pb}}\omega\,\langle \widehat{v}, \mathrm{sg}[h]\rangle, \tag{10}$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product and $\lambda_{\mathrm{pb}} > 0$ sets auxiliary strength. The subscript denotes pullback. Appendix B specifies suffix eligibility.

Constrained updates. For main gradient $g = \nabla_\theta \mathcal{L}_{\mathrm{flow}}$, project the auxiliary gradient onto spatial-LoRA coordinates, $g_{\mathrm{sp}} = \Pi_{\mathrm{sp}} \nabla_\theta \mathcal{L}_{\mathrm{pb}}$, and apply global and coordinate-wise budgets:

$$c_g = \min\left(1, \frac{\eta_g \|g\|_2}{\|g_{\mathrm{sp}}\|_2}\right), \qquad g'_{\mathrm{sp},j} = \mathrm{clip}(c_g g_{\mathrm{sp},j}, -\kappa|g_j|, \kappa|g_j|), \qquad g_{\mathrm{update}} = g + g'_{\mathrm{sp}}, \tag{11}$$

Here j indexes parameter coordinates, clip clamps to the specified interval, $c_g = 1$ when $g_{\mathrm{sp}} = 0$, $\eta_g = 0.005$, and $\kappa = 0.05$. Both budgets precede optimizer processing. The action projections receive $\nabla_\phi \mathcal{L}_{\mathrm{flow}}$; both objectives use the final action-corrected prediction.

Joint optimization and inference. Teacher-forced flow matching jointly trains action conditioning, CReSM, selected cross-attention parameters, and spatial LoRA; spatial distillation adds the constrained auxiliary update. The text encoder, VAE, and teacher remain frozen. Inference denoises each block with action correction and memory conditioning, then commits its clean state once; the spatial observer is used only during training.

4 EXPERIMENTS

We evaluate SurgWAVE on LiRAct and Cholec80 through benchmark comparisons, cumulative component analysis, qualitative examples, and expert assessment.

4.1 EXPERIMENTAL SETUP

Datasets and task. LiRAct contains 3,333 liver-resection episodes from 18 patients: 2,359 paired-view and 974 monocular episodes yield 5,692 camera-view records. Each input uses one view, an initial observation, a complete caption, and its extracted action descriptions (Section 3.1). Annotations identify instruments and activities, with tissue targets and image locations where available, without frame-level timing. Evaluation uses 1,310 records from 799 episodes and five patients excluded from adaptation. All task-specific training runs on LiRAct, including the component studies, use the same 1,943 videos from 13 training patients. Appendix A details the training and evaluation partitions.

Cholec80 (Twinanda et al., 2017) contains 80 cholecystectomy videos recorded at 25 fps, with phase and tool-presence annotations. We evaluate 237 held-out clips using image and caption conditioning. Appendix C specifies model-specific adaptation and input settings.

Implementation and metrics. SurgWAVE uses Wan2.2-TI2V-5B (Team Wan et al., 2025; Wan Team, 2025) with the LongLive-2.0 block-causal architecture (Chen et al., 2026b) and the joint optimization in Section 3.4. Tables 7 and 8 summarize training and inference settings. We report Fréchet Video Distance (FVD) (Unterthiner et al., 2018) and six VBench dimensions (Huang et al., 2024): subject and background consistency, imaging quality, motion smoothness, dynamic degree, and temporal flickering. FVD uses I3D features from 16-frame windows at 256-pixel resolution for LiRAct and 128 pixels for Cholec80, and is compared within each dataset. Dynamic degree measures detected motion, not action correctness; flickering uses eligible static clips. Appendix C details sampling and eligibility. The reference video represents one possible continuation.

4.2 BENCHMARK COMPARISONS

Compared methods. We evaluate Endora (Li et al., 2024), MOFA-Video (Niu et al., 2024), text-adapted SG2VID (Sivakumar et al., 2026), LingBot-World (Robbyant Team et al., 2026), Yume-1.5 (Mao et al., 2025), Cosmos-H-Surgical (He et al., 2025), and Ophora (Li et al., 2025). These system comparisons retain native input interfaces: Endora is unconditional, SG2VID uses text without scene graphs, and MOFA-Video receives reference-derived future flow. Output durations, resolutions, adaptation, and coverage are documented in Appendix C.

LiRAct results. SurgWAVE achieves FVD 154.37, compared with 260.71 for LingBot-World and 491.54 for Cosmos-H-Surgical: reductions of 106.34 (40.79%) and 337.17 (68.59%), respectively. Relative reductions use $(\mathrm{FVD}_{\mathrm{baseline}} - \mathrm{FVD}_{\mathrm{ours}})/\mathrm{FVD}_{\mathrm{baseline}}$. SurgWAVE leads in subject consistency, imaging quality, and motion smoothness, exceeding Cosmos-H-Surgical by 0.0288, 0.0546, and 0.0180. Its background consistency is below MOFA-Video's (0.9449 versus 0.9515). LingBot's comparison also reflects lower output coverage: 1,029 videos versus 1,310 for other methods.

Cholec80 results. SurgWAVE's FVD of 303.80 improves on Endora's next-lowest 633.78 by 329.98 (52.07%), and on Cosmos-H-Surgical by 379.02 (55.51%). Subject consistency, background consistency, imaging quality, and motion smoothness exceed the strongest alternative in each dimension by 0.0333, 0.0023, 0.0497, and 0.0068.

Table 1: LiRAct benchmark. Each method has 1,310 generated outputs, except LingBot-World-14B (1,029). Lower FVD is better; the VBench dimensions are defined in the text. Methods retain their listed conditioning interfaces (Appendix C); MOFA-Video uses reference-derived optical flow. N/A denotes an unavailable temporal-flickering estimate.

Method	FVD ↓	VBench					
		Subject	Background	Imaging	Motion	Dynamic	Flicker
Endora	2791.64	0.7101	0.8984	0.5886	0.9227	0.3443	0.9190
MOFA-Video	772.06	0.8643	**0.9515**	0.4435	0.9637	0.9718	0.9772
SG2VID (text-adapted)	1491.56	0.8732	0.9472	0.4281	0.9840	0.3168	0.9825
LingBot-World-14B	260.71	0.8804	0.9415	0.5766	0.9651	**1.0000**	N/A
Yume-1.5	1124.67	0.8613	0.9375	0.5024	0.9775	0.9298	0.9788
Cosmos-H-Surgical	491.54	0.8905	0.9505	0.5401	0.9804	0.9863	0.9839
SurgWAVE (ours)	**154.37**	**0.9193**	0.9449	**0.5947**	**0.9984**	**1.0000**	N/A

Table 2: Cholec80 benchmark on 237 held-out clips. FVD is computed from 16-frame windows at 128-pixel evaluation resolution; VBench evaluates the corresponding generated clips. N/A denotes an unavailable temporal-flickering estimate.

Method	FVD ↓	VBench					
		Subject	Background	Imaging	Motion	Dynamic	Flicker
Endora	633.78	0.7959	0.8953	0.5010	0.9595	0.7173	0.9627
Cosmos-H-Surgical	682.82	0.9013	0.9689	0.6515	0.9571	0.9873	0.9853
Yume-1.5	709.31	0.7644	0.8725	0.6230	0.9224	**1.0000**	N/A
Ophora	928.63	0.8498	0.9455	0.6085	0.9158	**1.0000**	N/A
SurgWAVE (ours)	**303.80**	**0.9346**	**0.9712**	**0.7012**	**0.9663**	**1.0000**	N/A

4.3 Component analysis

Study design. Table 3 evaluates cumulative configurations on the same LiRAct manifest and future-frame scoring rule. CKPT, native continuation (Cont.), and action conditioning (A) use 0, 1,000, and 1,000 additional updates. Cont. restarts its data loader after 300 updates, matching the total budget but not the exact sample order. Continuation reduces FVD from 392.80 to 337.93, a decrease of 54.87 (13.97%). Subsequent increments are conditional on the preceding configuration, rather than independent estimates of all component interactions.

Correlation-aware signed action conditioning. Adding A lowers FVD by 103.40 (30.60%), from 337.93 to 234.53. Subject consistency rises from 0.8464 to 0.8869 and motion smoothness from 0.9622 to 0.9706, while imaging quality decreases slightly from 0.5213 to 0.5193. This comparison assesses action conditioning as a whole. Figure 3 separately illustrates Independent aggregation, which omits cross-description Gram interactions, and Correlated aggregation using Eq. (5). Under the same ultrasonic-scalpel description, Correlated shows a more pronounced tissue opening at later times. The first three rows illustrate responses associated with transection, displacement, and traction. These examples visualize the mechanism without establishing a cohort-level correlation effect. Appendix D defines the configurations; Figure 7 provides further paired examples.

Predicted-state spatial distillation. Adding D reduces FVD by 18.80 (8.02%), from 234.53 to 215.73, while subject consistency, imaging quality, and motion smoothness rise by 0.0230, 0.0508, and 0.0285. These gains are consistent with local structural supervision, although the metrics do not directly measure anatomical correctness. The frozen observer adds no inference computation.

Contractive Residual State Memory (CReSM). Adding M lowers FVD from 215.73 to 154.37, a reduction of 61.36 (28.44%). Subject consistency rises from 0.9099 to 0.9193 and imaging quality from 0.5701 to 0.5947; motion smoothness changes from 0.9991 to 0.9984. The full model improves FVD over Cont. by 183.56 (54.32%). These aggregate results support the memory-enabled

Table 3: Cumulative component study on LiRAct. CKPT is the starting checkpoint and Cont. is native continuation with additional training. A, D, and M denote correlation-aware signed action conditioning, predicted-state spatial distillation, and CReSM. Each successive difference is conditional on the preceding configuration.

Configuration	FVD ↓	Subject ↑	Imaging ↑	Motion ↑
CKPT	392.80	0.8219	0.5345	0.9701
Cont.	337.93	0.8464	0.5213	0.9622
A	234.53	0.8869	0.5193	0.9706
$A + D$	215.73	0.9099	0.5701	**0.9991**
$A + D + M$	**154.37**	**0.9193**	**0.5947**	0.9984

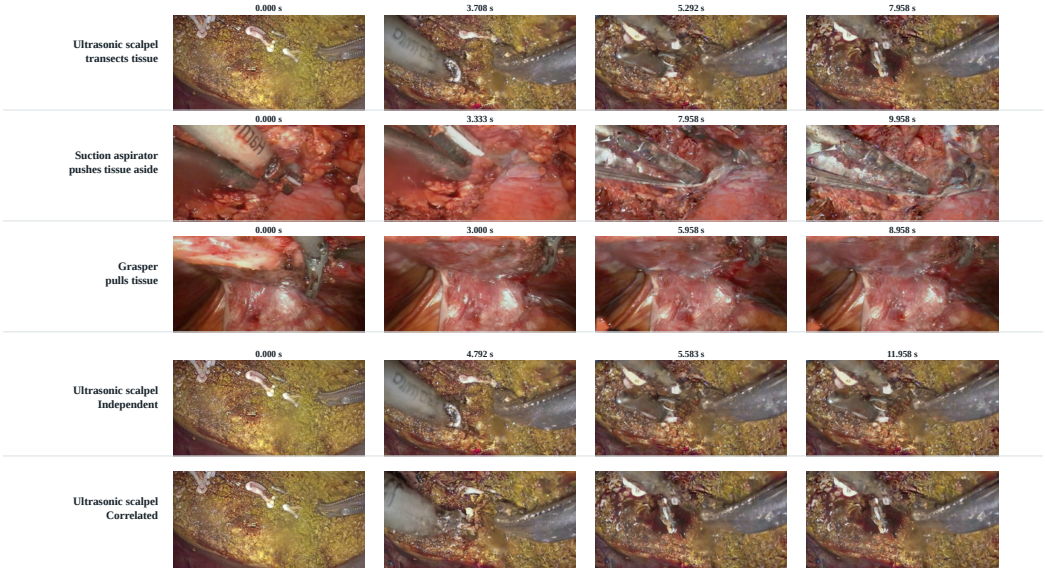

Figure 3: **Action-conditioned visual evolution.** The first three rows illustrate ultrasonic-scalpel transection, suction-aspirator tissue displacement, and grasper traction. The last two rows compare Independent and Correlated aggregation for the same ultrasonic-scalpel description at matched time points. Independent omits cross-description Gram interactions; Correlated uses Eq. (5).

configuration without isolating retention beyond the attention window. Reference-based evaluation extends to 15 seconds (Appendix C).

4.4 Qualitative evaluation

Figure 4 compares continuations involving right-angle forceps and suction. Yume-1.5 exhibits appearance drift and a late red color shift; LingBot-World preserves background texture but enlarges the foreground instrument shaft. SurgWAVE maintains more distinct shafts and a curved tip, while following a different motion trajectory from the recorded continuation. The example illustrates complementary aspects of instrument integrity and temporal coherence. Appendix H provides another case using the final SurgWAVE model.

4.5 Expert evaluation

Following structured surgical video assessment (Chen et al., 2025), four hepatobiliary surgeons from a major hospital, each with 5–10 years of surgical experience, anonymously assessed all methods on the same 80 held-out cases. Every surgeon rated every case for each method on four five-point scales. Event support averages the supported fraction of assessable actions over the 320 surgeon–case pairs per method (Appendix F).

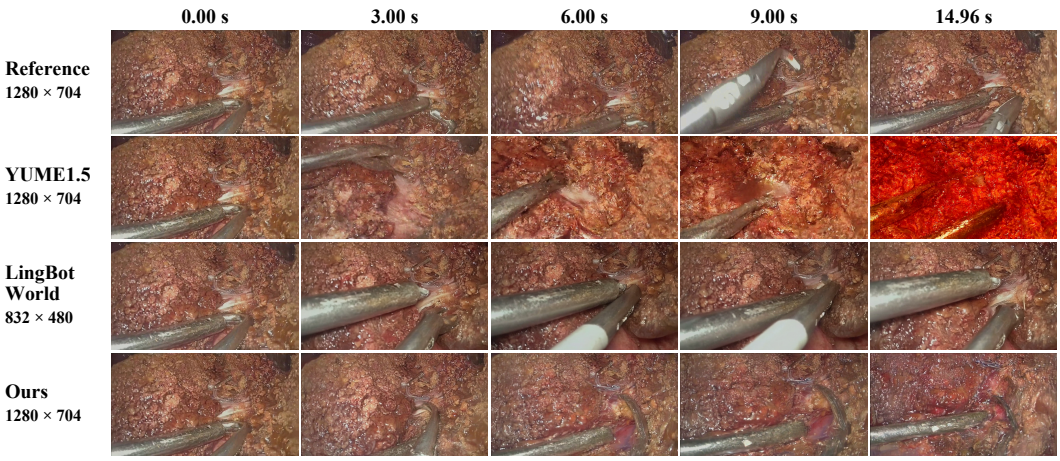

Figure 4: Qualitative comparison on LiRAct at aligned time points. Ours denotes the final SurgWAVE joint model reported in Table 1; Reference is the recorded future. Labels give native output resolutions, and aspect ratios are preserved. The reference is one observed continuation compatible with the untimed action descriptions.

Table 4: Anonymous expert evaluation on the same 80 cases per method. Activity, Instrument, Interaction, and Continuity denote mean five-point ratings of activity compatibility, instrument integrity, tool–tissue interaction, and temporal continuity. Event support is the mean of surgeon–case support fractions, expressed as a percentage (Eq. (12)).

Model	Activity	Instrument	Interaction	Continuity	Event support (%)
LingBot-World	3.38	3.50	2.75	4.63	63.2
Yume-1.5	3.00	3.25	2.50	4.50	61.5
SurgWAVE	**3.63**	**3.69**	**3.13**	**4.74**	**87.0**

SurgWAVE leads on all four rating dimensions (Table 4). Its activity, instrument, interaction, and continuity scores exceed LingBot-World's by 0.25, 0.19, 0.38, and 0.11 points, and Yume-1.5's by 0.63, 0.44, 0.63, and 0.24. Event support reaches 87.0%, gains of 23.8 and 25.5 percentage points, respectively. Larger differences in event support and interaction complement smaller continuity gains, showing why smoothness alone does not characterize activity compatibility. These are descriptive comparisons without uncertainty estimates.

5 Conclusion

SurgWAVE combines correlation-aware signed action conditioning, Contractive Residual State Memory (CReSM), and predicted-state spatial distillation for surgical video generation from untimed descriptions. We introduce LiRAct to evaluate this setting. SurgWAVE achieves the lowest reported FVD among the evaluated methods on LiRAct and Cholec80 and higher descriptive expert ratings on shared LiRAct cases. Current reference clips extend to 15 seconds; evaluating substantially longer surgical rollouts remains future work.

AI use statement

Generative AI tools were used as assistive tools for coding, literature retrieval, drafting of the manuscript. This assistance included feedback on mathematical exposition and interpretation of reported results. All AI-assisted material was subsequently reviewed and further revised by the authors. The authors take full responsibility for the final manuscript, including its methods, analyses, and conclusions.

Reproducibility statement

The Method and Experiments sections describe the model formulation, assumptions, training objectives, and evaluation settings. The appendix provides supporting derivations, data partitions, reported training settings, and evaluation protocols, including the expert-rating procedure. The code and benchmark will be released upon acceptance.

References

Tong Chen, Shuya Yang, Junyi Wang, Long Bai, Hongliang Ren, and Luping Zhou. SurgSora: Object-aware diffusion model for controllable surgical video generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15969 of *Lecture Notes in Computer Science*, pp. 521–531. Springer Nature Switzerland, 2026a. doi: 10.1007/978-3-032-05127-1_50. URL https://link.springer.com/chapter/10.1007/978-3-032-05127-1_50.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, Yuyang Zhao, Huizi Mao, Ying-Cong Chen, Enze Xie, Xiaojuan Qi, and Song Han. LongLive-2.0: An NVFP4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026b. doi: 10.48550/arXiv.2605.18739. URL https://arxiv.org/abs/2605.18739.

Zhen Chen, Qing Xu, Jinlin Wu, Biao Yang, Yuhao Zhai, Geng Guo, Jing Zhang, Yinlu Ding, Nassir Navab, and Jiebo Luo. How far are surgeons from surgical world models? a pilot study on zero-shot surgical video generation with expert assessment. *arXiv preprint arXiv:2511.01775*, 2025. doi: 10.48550/arXiv.2511.01775. URL https://arxiv.org/abs/2511.01775.

Michaël Fanuel, Joachim Schreurs, and Johan A. K. Suykens. Diversity sampling is an implicit regularization for kernel methods. *SIAM Journal on Mathematics of Data Science*, 3(1):280–297, 2021. doi: 10.1137/20M1320031. URL https://doi.org/10.1137/20M1320031.

Yang Fei, George Stoica, Jingyuan Liu, Qifeng Chen, Ranjay Krishna, Xiaojuan Wang, and Benlin Liu. Structure from tracking: Distilling structure-preserving motion for video generation. *arXiv preprint arXiv:2512.11792*, 2025. URL https://arxiv.org/abs/2512.11792.

Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception. *arXiv preprint arXiv:2607.05247*, 2026.

Marta Garnelo and Wojciech Marian Czarnecki. Exploring the space of key-value-query models with Intention. *arXiv preprint arXiv:2305.10203*, 2023. doi: 10.48550/arXiv.2305.10203. URL https://arxiv.org/abs/2305.10203.

Yufan He, Pengfei Guo, Mengya Xu, Zhaoshuo Li, Andriy Myronenko, Dillan Imans, Bingjie Liu, Dongren Yang, Mingxue Gu, Yongnan Ji, Yueming Jin, Ren Zhao, Baiyong Shen, and Daguang Xu. Cosmos-H-Surgical: Learning surgical robot policies from videos via world modeling. *arXiv preprint arXiv:2512.23162*, 2025. doi: 10.48550/arXiv.2512.23162. URL https://arxiv.org/abs/2512.23162.

Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, Yingnian Wu, and Lijuan Wang. SlowFast-VGen: Slow-fast learning for action-driven long video generation. In *International Conference on Learning Representations*, pp. 76609–76636, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/beb743f699ce0fc9e86f786b61f0c815-Abstract-Conference.html.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. LongLive-RAG: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026. doi: 10.48550/arXiv.2606.02553. URL https://arxiv.org/abs/2606.02553.

Jiehui Huang, Yuechen Zhang, Xu He, Yuan Gao, Zhi Cen, Bin Xia, Yan Zhou, Xin Tao, Pengfei Wan, and Jiaya Jia. UnityVideo: Unified multi-modal multi-task learning for enhancing world-aware video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2512.07831.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-free infinite video generation via evolving memory tokens. *arXiv preprint arXiv:2603.12513*, 2026. doi: 10.48550/arXiv.2603.12513. URL https://arxiv.org/abs/2603.12513.

Bohan Li, Shuojue Yang, Baorui Peng, Xianda Guo, Erli Zhang, Youqi Tao, Junfeng Duan, Daguang Xu, Qi Dou, Xin Jin, Wenjun Zeng, Hao Zhao, and Yueming Jin. From articulated kinematics to routed visual control for action-conditioned surgical video generation. *arXiv preprint arXiv:2605.08712*, 2026. doi: 10.48550/arXiv.2605.08712. URL https://arxiv.org/abs/2605.08712.

Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y. Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006 of *Lecture Notes in Computer Science*, pp. 230–240. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72089-5_22. URL https://papers.miccai.org/miccai-2024/275-Paper0470.html.

Wei Li, Ming Hu, Guoan Wang, Lihao Liu, Kaijing Zhou, Junzhi Ning, Xin Guo, Zongyuan Ge, Lixu Gu, and Junjun He. Ophora: A large-scale data-driven text-guided ophthalmic surgical video generation model. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15968 of *Lecture Notes in Computer Science*, pp. 425–435. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05114-1_41. URL https://papers.miccai.org/miccai-2025/0656-Paper1768.html.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-Adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ny8T8OuNHe.

Huanrong Liu, Weiliang Huang, Bob Zhang, Weichao Cai, Chunlin Tian, and Qingbiao Li. SurgWMBench: A vision-based benchmark for world-modeling surgical instrument motion planning. *arXiv preprint arXiv:2608.08070*, 2026. doi: 10.48550/arXiv.2608.08070. URL https://arxiv.org/abs/2608.08070.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. doi: 10.48550/arXiv.2509.25161. URL https://arxiv.org/abs/2509.25161.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025. doi: 10.48550/arXiv.2512.22096. URL https://arxiv.org/abs/2512.22096.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. MOFA-Video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *Computer Vision – ECCV 2024*, volume 15077 of *Lecture Notes in Computer Science*, pp. 111–128. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72655-2_7. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/2842_ECCV_2024_paper.php.

Wentao Pan, Wuyang Li, Shengyuan Liu, Xinyu Liu, Hengyu Liu, and Yixuan Yuan. SurgVista: Long-horizon surgical world modeling with plausible instrument-tissue dynamics. *arXiv preprint arXiv:2606.19889*, 2026. doi: 10.48550/arXiv.2606.19889. URL https://arxiv.org/abs/2606.19889.

Sampath Rapuri, Lalithkumar Seenivasan, Dominik Schneider, Roger Soberanis-Mukul, Yufan He, Hao Ding, Jiru Xu, Chenhao Yu, Chenyan Jing, Pengfei Guo, Daguang Xu, and Mathias Unberath. SAW: Toward a surgical action world model via controllable and scalable video generation. *arXiv preprint arXiv:2603.13024*, 2026. doi: 10.48550/arXiv.2603.13024. URL https://arxiv.org/abs/2603.13024.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. doi: 10.48550/arXiv.2601.20540. URL https://arxiv.org/abs/2601.20540.

Ssharvien Kumar Sivakumar, Yannik Frisch, Ghazal Ghazaei, and Anirban Mukhopadhyay. SG2VID: Scene graphs enable fine-grained control for video synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15968 of *Lecture Notes in Computer Science*, pp. 511–521. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-05114-1_49. URL https://link.springer.com/chapter/10.1007/978-3-032-05114-1_49.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. doi: 10.48550/arXiv.2503.20314. URL https://arxiv.org/abs/2503.20314.

Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017. doi: 10.1109/TMI.2016.2593957. URL https://doi.org/10.1109/TMI.2016.2593957.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. doi: 10.48550/arXiv.1812.01717. URL https://arxiv.org/abs/1812.01717.

Wan Team. Wan2.2: Official model and code release. https://github.com/Wan-Video/Wan2.2, 2025. TI2V-5B release; accessed September 15, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations*, pp. 90037–90059, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/91a1610c6ed9e02d33f826b46f472b92-Abstract-Conference.html.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8319–8329. IEEE, 2025.

Guy Yariv, Yuval Kirstain, Amit Zohar, Shelly Sheynin, Yaniv Taigman, Yossi Adi, Sagie Benaim, and Adam Polyak. Through-The-Mask: Mask-based motion trajectories for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 18198–18208, 2025. URL https://arxiv.org/abs/2501.03059.

Yousef Yeganeh, Rachmadio Lazuardi, Amir Shamseddin, Emine Dari, Yash Thirani, Nassir Navab, and Azade Farshad. VISAGE: Video synthesis using action graphs for surgery. *arXiv preprint arXiv:2410.17751*, 2024. doi: 10.48550/arXiv.2410.17751. URL https://arxiv.org/abs/2410.17751.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, and Xiaojuan Qi. VideoSSM: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025. doi: 10.48550/arXiv.2512.04519. URL https://arxiv.org/abs/2512.04519.

Lin Zhao, Yushu Wu, Yifan Gong, Yanzhi Wang, and Pu Zhao. OmniMem: Scalable and adaptive memory retrieval for long video generation. *arXiv preprint arXiv:2605.30519*, 2026. doi: 10.48550/arXiv.2605.30519. URL https://arxiv.org/abs/2605.30519.

Qiang Zhou, Shaofeng Zhang, Nianzu Yang, Ye Qian, and Hao Li. Motion control for enhanced complex action video generation. *arXiv preprint arXiv:2411.08328*, 2024. URL https://arxiv.org/abs/2411.08328.

Rulin Zhou, Wanhao Liu, Guoheng Ma, Liangjin Shao, Qiujie Song, Yidu Wang, Guankun Wang, Tong Chen, Long Bai, Luping Zhou, and Hongliang Ren. Surg-UniWorld: A unified surgical world model with multimodal control experts. *arXiv preprint arXiv:2608.06770*, 2026. doi: 10.48550/arXiv.2608.06770. URL https://arxiv.org/abs/2608.06770.

A Dataset and annotations

The supplementary material records the data split, model configuration, generation protocol, component configurations, and expert evaluation procedure used throughout the paper.

Episodes and camera views. An operative episode denotes a continuous surgical segment, whereas a camera-view record corresponds to a single view of that episode. LiRAct comprises 5,692 records from 18 patients: 2,359 paired-view episodes and 974 monocular episodes. Episode identifiers encode the patient, source recording, annotation table, and clip index. Paired views share the patient, data partition, caption, and temporal boundaries. Each generator input contains a single camera view.

Table 5: LiRAct source collection and patient split before preparation. The shared training set and evaluation cohorts are listed in Table 6.

Unit	All	Original training partition	Original evaluation partition
Patients	18	13	5
Physical episodes	3,333	2,505	828
View-level records	5,692	4,339	1,353

Training and evaluation partitions. All task-specific training runs for the LiRAct experiments, including the component studies, use the same training set of 1,943 videos from 13 patients (Table 6). The 1,310-record evaluation set contains 799 physical episodes from five held-out patients: 511 paired-view episodes and 288 monocular episodes. Caption extraction preserves wording and repetition; all selected captions are parseable under the annotation template.

Table 6: Shared training set and evaluation cohorts. Counts distinguish training videos, evaluation records or clips, and expert-evaluation cases.

Subset	Patients	Count	Counted unit
Shared LiRAct training set	13	1,943	Videos
Full LiRAct evaluation	5	1,310	View-level records
Cholec80 evaluation	–	237	Held-out clips
Expert evaluation	5	80	Shared video cases

The five evaluation patients are excluded from the shared training set. Preprocessing reduces the original 1,353 evaluation records to 1,310 prepared records; the same manifest is used for all matched comparisons.

Annotation semantics. Each action sentence specifies an instrument, an activity, and, when available, a target; examples include *Grasper retracts tissue* and *Dissector separates tissue*. Multiple activities may occur within a segment. Repeated sentences are retained as distinct multiset entries, with their original order preserved in the complete caption. These annotations provide segment-level activity descriptions rather than event timestamps or occurrence counts. Generation uses both the complete caption and its extracted action descriptions.

Evaluation manifests use anonymous identifiers and fixed patient-level partitions. Paired camera views of an episode remain in the same partition.

B Architecture and training

Backbone and signed action-conditioning architecture. The generator comprises 30 transformer blocks with hidden width 3,072 and 24 attention heads of width 128. Its VAE encodes 48 latent channels. Action sentences are extracted according to the standardized annotation template, while the complete caption is retained in the backbone's text pathway. Valid T5 token features are mean-pooled separately for each sentence to obtain a 4,096-dimensional embedding. A shared LayerNorm, $4096 \to 128$ projection, and GELU feed distinct 64-dimensional key and 48-dimensional

signed-value heads. The query projection maps the concatenation of the noisy latent, native flow v_θ^{native}, and noise level to 64 dimensions. The value head is zero-initialized. The action-conditioning projections contain 553,328 trainable parameters, and the regularized solve is computed in FP32. Repeated sentences share encoder computation but retain their multiplicity in the action multiset.

Contractive Residual State Memory (CReSM). CReSM augments four self-attention layers with an eight-dimensional state per head. The encoder, decoder, and retention parameters contribute 196,612 parameters. Retention is parameterized by $\rho = \rho_{\max} \operatorname{sigmoid}(r)$, where $r \in \mathbb{R}$ is a learned retention logit and $0 < \rho_{\max} < 1$ is its fixed upper bound; the decoder is zero-initialized. Anchors and committed states are stored in FP32 separately for the two classifier-free-guidance branches. The joint model applies the full bounded residual in Eq. (8). Appendix G specifies the corresponding state transitions.

Predicted-state spatial observer. The frozen LingBot-Vision teacher is ViT-small with patch size 16 and four register tokens. The observer decodes a four-latent suffix and processes its final RGB frame. The image is mapped to $[0, 1]$, resized bilinearly to 384×384, and normalized with the teacher's ImageNet statistics. Final normalized patch tokens, excluding register tokens, form a 24×24 grid and are L2-normalized in FP32. Excluding one border patch leaves a 22×22 grid, with $2 \times 22 \times 21 = 924$ horizontal and vertical cosine-dissimilarity relations.

An eligible suffix contains four valid latent frames with a common task-loss weight and scheduler index s satisfying $0 < s \leq s_{\text{obs}}$, where s_{obs} selects the low-noise range. The first eligible block suffix is used within each local temporal shard; if none is eligible, the auxiliary contribution is zero. The scheduler maps discrete indices to the noise scale σ. Rademacher probes estimate the trace normalization, with a positive floor. Both the detached clean reference and the flow error are computed from the action-corrected prediction $\widehat{v}$. The auxiliary backward contributes only to the spatial LoRA parameters, while the main flow objective updates the action-conditioning projections, CReSM, selected cross-attention parameters, and spatial LoRA. The text encoder, VAE, and LingBot observer remain frozen.

B.1 Training configuration

Table 7 summarizes the main training configuration. Optimization follows Section 3.4, with the spatial auxiliary update restricted to spatial LoRA. Inference settings are reported separately in Table 8.

Table 7: Summary of the main training configuration.

Setting	Configuration
Backbone initialization	Wan2.2-TI2V-5B with the LongLive-2.0 block-causal architecture
Main objective	Teacher-forced flow matching
Trainable components	Action-conditioning projections, CReSM, selected cross-attention parameters, and spatial LoRA
Frozen components	Text encoder, VAE, and LingBot-Vision teacher
Learning rate	3×10^{-5}
Learning-rate schedule	Warm-up followed by cosine decay
Batch configuration	Global batch size 8 across 8 GPUs; 1 sample per GPU

C Generation and evaluation protocols

Length and history. Inference allocates a multiple of eight latent frames, clamps the initial image in the first output slot, and crops decoded padding to the manifest's retained RGB length. The 1,310-record evaluation manifest has 23–360 retained RGB frames per record, with median 263.5; at 24 fps the maximum duration is 15 seconds. The rolling cache stores $3 \times 8 \times 880 = 21{,}120$ tokens per layer, corresponding to 24 latent frames. This cache defines the available local attention context.

Table 9 describes available reference duration, not a separate horizon-stratified performance experiment. The reported reference-based comparisons use the available recorded continuations; they do not establish performance at 30-second, minute-scale, or full-procedure horizons.

Table 8: LiRAct inference configuration for the module studies. Variant-specific context settings are identified explicitly. External-baseline settings are described below.

Setting	Value
Output resolution (width × height)	1280×704
Frame rate	24 fps
Denoising steps	50
Classifier-free guidance scale	3
Timestep shift	5
Generation seed / samples per condition	3407 / 1
Block size / rolling cache	8 / 24 latent frames
Attention tokens per latent frame	880
Action/native configuration sink	0 latent frames
CReSM configuration sink	1 latent frame
Action configuration RoPE offset	8

Table 9: Available durations in the 1,310-record evaluation set. Frame counts include the initial observation and exclude padding.

Retained duration	Minimum RGB frames	Available records
4 seconds	96	1,169
8 seconds	192	902
12 seconds	288	597
16 seconds	384	0

Automatic measures. FVD is a dataset-level distance between I3D video features. We extract features from 16-frame windows at 256-pixel evaluation resolution for LiRAct and 128-pixel evaluation resolution for Cholec80, using reproducible filename- and seed-based window selection. FVD values are compared within each dataset. VBench subject and background consistency use whole-frame DINO and CLIP similarities. Imaging quality uses a learned image-quality assessor; motion smoothness measures temporal interpolation behavior; dynamic degree measures the fraction of videos classified as moving by optical-flow thresholds; and temporal flickering is evaluated on eligible static clips. Output coverage and the static-video counts available for temporal flickering are listed in Table 10. N/A in a result table denotes an unavailable score and does not establish that the eligible subset is empty or that flickering is absent. These measures describe visual and temporal properties; expert assessment evaluates surgical activity and tool–tissue interaction.

Controlled comparisons use observed-context and padded-frame exclusion, fixed evaluator versions, and method-specific valid-output accounting. Reference-based scores measure similarity to the recorded continuation. Table 10 lists baseline checkpoints, input controls, and available output counts; each metric uses its eligible subset.

Endora, MOFA-Video, and SG2VID were evaluated on 128×128 outputs of 16, 25, and 16 frames, respectively. MOFA-Video receives optical flow extracted from the recorded future; the SG2VID adapter uses image and text without scene graphs. LingBot outputs are 832×480, while Yume-1.5 and SurgWAVE outputs are 1280×704. Their exported playback rate is 24 fps; LingBot and Yume generate natively at 16 fps. LingBot's listed FVD uses its 1,029 available outputs. The benchmark uses the stated evaluation cohorts and preprocessing, while retaining the method-specific conditioning interfaces, output durations, and available output counts documented here. These system-level comparisons should therefore be distinguished from the matched component comparisons.

LingBot-World inference. The baseline uses the Base-Cam model release.[1] Evaluation uses image and text without camera trajectories or keyboard controls, with UniPC, 70 denoising steps, shift 10, and guidance 5. LingBot-Vision serves as the distinct spatial teacher in Section 3.4.

Statistical interpretation. The tables report descriptive point estimates. Repeated frames, paired views of an episode, and repeated clinical ratings are correlated observations. No uncertainty es-

[1] `https://huggingface.co/robbyant/lingbot-world-base-cam`

Table 10: Evaluated configurations and inputs. Output N is the number of generated videos; Flicker N is the static-video subset used for temporal flickering. Dataset sections correspond to Tables 1 and 2.

Method	Checkpoint / adaptation	Input controls	Output N	Flicker N
LiRAct				
Endora	Kvasir initialization; 1,000 updates	Unconditional	1,310	859
MOFA-Video	–	Image and reference-derived flow	1,310	37
SG2VID adapter	–	Image and text; no scene graph	1,310	895
LingBot-World	Base-Cam release	Image and text; no action trajectory	1,029	0
Yume-1.5	Shared-core adaptation; 1,000 updates	Image and text; fixed-camera prefix	1,310	92
Cosmos-H-Surgical	Predict2.5-2B surgical release; 1,000 updates	Image and text	1,310	18
SurgWAVE	Joint model; 1,000 updates	Image, caption and action descriptions	1,310	2
Cholec80				
Endora	Kvasir initialization; 10,000 updates	Unconditional	237	67
Cosmos-H-Surgical	Predict2.5-2B surgical release; 1,000 updates	Image and caption	237	3
Yume-1.5	Shared-core adaptation; 1,000 updates	Image and caption	237	0
Ophora	Released Ophora model; 1,000 updates	Caption only	237	0
SurgWAVE	Joint model; 1,000 updates	Image and caption	237	N/A

timates are reported, so numerical differences should not be interpreted as tests of statistical significance. Method-specific input controls and output coverage are retained in cross-architecture comparisons.

D COMPONENT CONFIGURATIONS

Table 3 reports cumulative configurations: CKPT is the starting checkpoint, Cont. is native continuation, A adds correlation-aware signed action conditioning, D adds predicted-state spatial distillation, and M adds CReSM. Each increment is conditional on the preceding configuration and does not isolate all interactions. The training-budget and data-order qualifications are given in Section 4.3.

The qualitative Independent control omits cross-description Gram interactions, whereas Correlated uses Eq. (5). Figures 3 and 7 show these configurations under the same supplied descriptions. No separate cohort-level Independent-versus-Correlated estimate is reported.

E SUPPLEMENTARY ACTION DIAGNOSTICS

Figure 5 presents the action-aggregation diagnostics, including reorder behavior, correction strength, case composition, and event support. Figure 6 provides an illustrative reference for the four diagnostic perspectives.

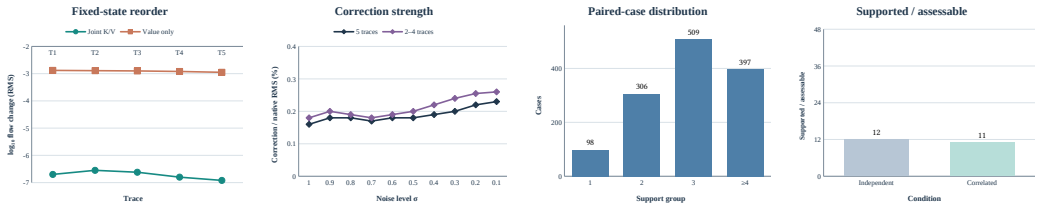

Figure 5: Supplementary action-aggregation diagnostics. The four panels show fixed-state reorder, correction strength, paired-case distribution, and supported-versus-assessable outcomes.

F EXPERT EVALUATION PROTOCOL

Cases and assessors. All evaluated methods use the same 80 video cases from the held-out LiRAct cohort. The anonymous assessment was conducted by four hepatobiliary surgeons from a major hospital, each with 5–10 years of surgical experience. Every surgeon assessed all 80 cases for every evaluated method, yielding 320 surgeon–case assessments per method. Thus, 80 is the number of shared cases per method, not the total number of generated videos across methods. Each case is defined by its initial observation and supplied action description.

Rating dimensions. Surgeons assess activity compatibility, instrument integrity, tool–tissue interaction, and temporal continuity on five-point ordinal scales. Table 11 specifies the rating anchors.

Event support and aggregation. For method m, surgeon $r \in \{1, \ldots, 4\}$, and case $c \in \{1, \ldots, 80\}$, let $n_{mrc} > 0$ be the number of assessable annotated action events and k_{mrc} the number judged supported by the generated continuation. The surgeon–case support fraction is $p_{mrc} = k_{mrc}/n_{mrc}$. The reported event support is the unweighted mean over the four surgeons and 80 cases:

$$\mathrm{ES}_m = \frac{100}{4 \times 80} \sum_{r=1}^{4} \sum_{c=1}^{80} p_{mrc}. \tag{12}$$

Thus, support is computed within each surgeon–case pair and then averaged across these pairs; it is neither a pooled event proportion nor a count of supported videos divided by 80. Five-point ratings and event-support percentages are descriptive summaries without uncertainty estimates. They assess visible activity compatibility and surgical plausibility, rather than clinical safety or patient outcomes.

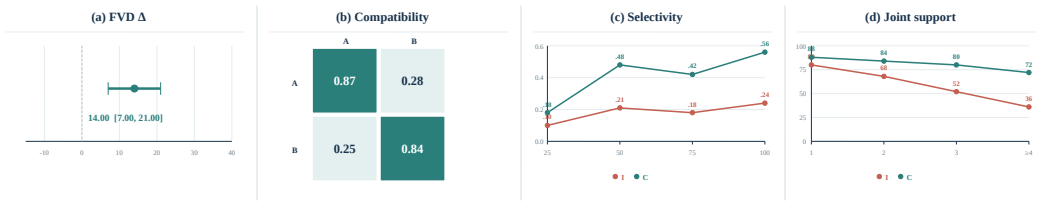

Figure 6: **Illustrative action diagnostics.** The four panels present idealized reference patterns for aggregation, action compatibility, selectivity, and joint support.

Table 11: Expert-rating rubric for the shared 80-case assessment. Scores 2 and 4 are intermediate values.

Dimension	Score 1	Score 3	Score 5
Activity compatibility	Visible behavior contradicts the activity	Partial or inconsistent support	Clear visible behavior compatible with the activity
Instrument integrity	Major duplication, disappearance, or deformation	Intermittent local defects	Coherent visible identity and shape
Interaction plausibility	Clearly implausible contact or tissue response	Mixed or locally questionable response	Consistently plausible visible interaction
Temporal continuity	Major jumps, repetition, or state reversals	Intermittent drift or discontinuity	Coherent evolution, including genuine changes

G MATHEMATICAL DETAILS

Closed-Form Derivation of the Action Aggregator. Suppress frame and spatial indices. Let $K \in \mathbb{R}^{d\times N}$ and $R \in \mathbb{R}^{C\times N}$ be the weighted key and value matrices, $q \in \mathbb{R}^d$ the current query, and $G = K^\top K$. Write $[N] = \{1, \dots, N\}$ and let $Z = \det(I_N + G/\lambda)$ be the subset-weight normalizer, with $\lambda > 0$. For $S \subseteq [N]$, the selection matrix $P_S \in \mathbb{R}^{N\times|S|}$ contains the columns of I_N indexed by S; thus $K_S = KP_S$, $R_S = RP_S$, and $G_{SS} = P_S^\top G P_S$. Here $|S|$ is the subset size and π is Eq. (4) with frame indices omitted. The principal-minor expansion gives

$$Z = \sum_{S\subseteq[N]} \lambda^{-|S|} \det(G_{SS}). \tag{13}$$

Taking $\frac{1}{2}\nabla_K \log Z$ yields

$$K(G + \lambda I_N)^{-1} = \mathbb{E}_{S\sim\pi}\left[K_S (G_{SS})^\dagger P_S^\top\right], \tag{14}$$

Transposing and multiplying by R and q gives Eq. (5) for arbitrary signed R. To include rank-deficient keys, observe that the Cauchy–Binet expansion expresses each Gram determinant as a sum of squared minors of K_S. At rank deficiency these minors vanish, so both the determinant and its derivative with respect to K are zero; when $|S| > d$, the determinant is identically zero. Such subsets therefore contribute neither probability nor derivative. The identity therefore holds for the expected prediction operator, including rank-deficient key matrices. Evaluation requires a regularized linear solve, without subset enumeration or explicit inversion (Fanuel et al., 2021).

For a permutation matrix $P \in \mathbb{R}^{N\times N}$, replacing (K, R) by (KP, RP) leaves the operator unchanged because $P^\top G P + \lambda I_N = P^\top (G + \lambda I_N) P$. This property holds for a fixed pre-action prediction and query; the complete caption retains its original text encoding. Duplicating a column changes the multiset and generally changes the regularized response. Equation (6) bounds the frame-wise correction RMS by $\eta\,\mathrm{RMS}(v_\theta^{\mathrm{native}})$ through its common post-aggregation scale.

Boundedness and conditional contraction of the memory state. Each innovation coordinate lies in $[-1, 1]$, so the convex recurrence from $m_0 = 0$ keeps every state coordinate in this interval. For

fixed ρ and a common innovation sequence, let m_b and m'_b be trajectories initialized at m_0 and m'_0. Then $\|m_b - m'_b\|_\infty \le \rho^b \|m_0 - m'_0\|_\infty$: the state recurrence is contractive with respect to its initial state. Its history expansion is $m_b = (1-\rho)\sum_{j=1}^{b} \rho^{b-j} u_j$. Persistent innovations are continually integrated; for $u_j = u$, this gives $m_b = (1-\rho^b)u$. The full-strength readout satisfies $\mathrm{RMS}(\mathcal{R}_t) \le \epsilon_m s_t$, with s_t including the numerical floor in Eq. (8). This conditional contraction statement does not establish contraction of the coupled generator, whose innovations depend on generated features.

Local spatial metric. Let n_x be the number of scalar latent coordinates in the selected suffix and n_r the number of retained patch relations ($n_r = 924$ in our configuration). The observer maps $\mathcal{O} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_r}$, so $J \in \mathbb{R}^{n_r \times n_x}$ and $e, h \in \mathbb{R}^{n_x}$. At the fixed predicted state, $J^\top J/\tau$ is positive semidefinite since $\tau > 0$. A Rademacher vector $\xi \in \{-1, +1\}^{n_r}$ with independent equiprobable signs satisfies $\mathbb{E}[\xi\xi^\top] = I_{n_r}$, giving $\mathbb{E}\|J^\top \xi\|_2^2 = \mathrm{tr}(J^\top J)$ (Hutchinson, 1989). This probe estimate is unbiased before flooring and supplies the scalar normalization; the numerator $J^\top Je$ is computed by Jacobian products. Holding J, τ, and ω fixed, the local quadratic

$$Q(e) = \frac{\lambda_{\mathrm{pb}}\omega}{\tau}\|Je\|_2^2, \qquad e = \widehat{v} - v^\star, \tag{15}$$

has derivative $2\lambda_{\mathrm{pb}}\omega J^\top Je/\tau$. The stopped surrogate in Eq. (10) produces the same output derivative at the reference prediction. Thus errors that induce larger changes in the teacher's local relations receive a stronger auxiliary constraint.

Let θ denote the trainable generator parameters and ϕ the action-conditioning parameters. The main gradient $g = \nabla_\theta \mathcal{L}_{\mathrm{flow}}$ includes CReSM, selected cross-attention, and spatial LoRA. The auxiliary vector g_{sp} retains only spatial LoRA coordinates and is zero elsewhere in θ. Global rescaling followed by coordinate clipping preserves both budgets in Eq. (11); the resulting vector is added to g before optimizer processing. The action-conditioning parameters are updated separately through $\nabla_\phi \mathcal{L}_{\mathrm{flow}}$. These bounds apply to the gradients before optimizer transformation.

H Computational cost and supplementary material

Storage and evaluation records. The action-conditioning branch and CReSM introduce 553,328 and 196,612 parameters, respectively. The anchor and committed states require $2 \times 4 \times 24 \times 8$ FP32 scalars, or 6,144 bytes per sample and classifier-free-guidance branch; this analytical storage cost excludes temporary tensors, the rolling KV cache, and allocator overhead. Each evaluation record associates an anonymous source identifier with the model variant, sampling configuration, evaluator version, and valid sample count for each metric.

Additional qualitative comparison. Figure 8 presents an eight-second continuation conditioned on an event description involving right-angle and bipolar forceps. The row labeled Ours in both qualitative comparisons is the final SurgWAVE joint model reported in the benchmark tables. LingBot outputs have resolution 832×480, compared with 1280×704 for the remaining rows; frames are displayed at equal width while preserving their original aspect ratios.

Paired action outcomes. Figure 7 collects paired real-data rollouts for three actions under matched inputs. Each action is shown at four aligned time points with Independent and Correlated aggregation in adjacent rows, so the action-specific evolution can be inspected without changing the surrounding scene.

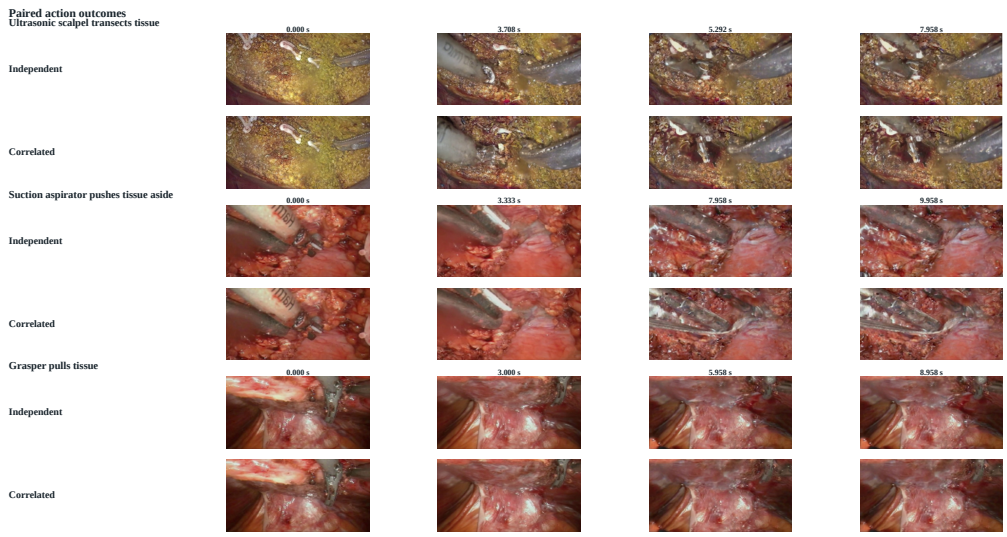

Figure 7: **Paired action outcomes.** Real-data paired rollouts for ultrasonic-scalpel transection, suction-aspirator tissue displacement, and grasper traction. For each action, Independent and Correlated aggregation are shown at aligned time points under the same input description.

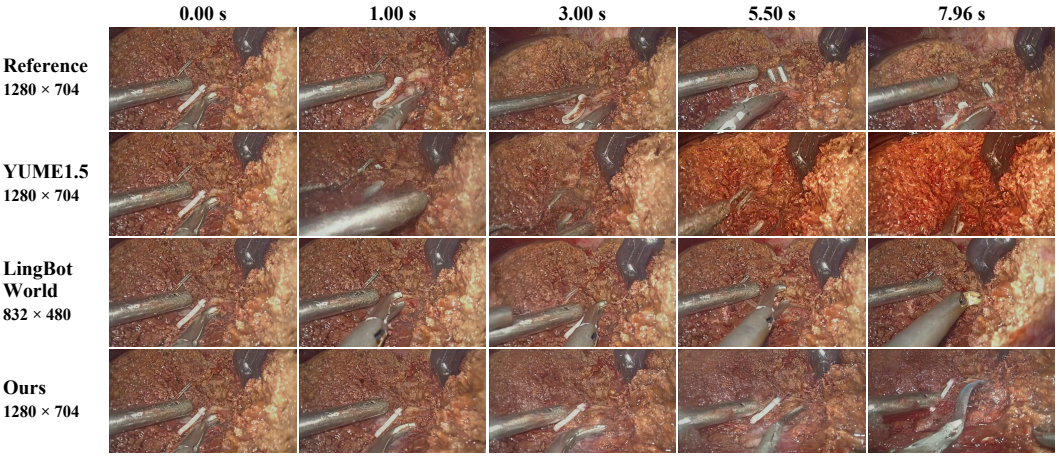

Figure 8: Additional qualitative comparison on LiRAct. Ours denotes the final SurgWAVE joint model. Row order and temporal sampling follow Figure 4. The recorded sequence provides a reference realization for the continuation.