CAUSAL VISUAL–TRAJECTORY JOINT PREDICTION FOR SURGICAL MOTION PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective surgical motion planning requires jointly anticipating future visual states and instrument trajectories, together with an explicit causal relationship between the predicted motion plan and the generated visual future. Existing surgical world models typically treat video generation and trajectory prediction as separate objectives or provide limited evidence that predicted trajectories meaningfully affect visual dynamics. To solve this problem, we propose a joint visual–trajectory prediction framework that establishes an explicit trajectory-to-visual pathway through spatially grounded motion conditioning. The framework transforms predicted trajectories into temporally aligned spatial representations and associates them with observed local appearance to guide future visual dynamics. We further introduce a visual–trajectory causal evaluation framework that measures local visual responses at spatial positions and time steps associated with instrument trajectories. Trajectory-Grounded Motion Alignment (TrajAlign) and Trajectory-Grounded Response Gain (TrajGain) characterize the directional consistency and relative strength of these responses, enabling quantitative assessment of how trajectory conditioning influences future visual dynamics. Experiments demonstrate improved trajectory accuracy and visual–trajectory consistency while maintaining competitive visual quality. Controlled response analysis provides quantitative evidence that trajectory conditioning induces directionally consistent changes in local visual motion, making the trajectory–visual relationship both explicit and measurable.

1 INTRODUCTION

Anticipating future instrument motion is fundamental to surgical motion planning. In surgery, however, an instrument trajectory is not meaningful in isolation: its execution is accompanied by localized tissue deformation, occlusion, appearance changes, and interactions with surrounding anatomy. A useful surgical world model should therefore predict not only where an instrument is likely to move, but also how the visual scene will evolve accordingly. More importantly, the predicted motion plan should have a causal relationship with the generated visual future, such that changes in the planned trajectory lead to corresponding changes in future visual dynamics. Jointly modeling these two futures provides a unified basis for forecasting both the intended motion and its potential visual consequences.

Motivated by this goal, recent approaches toward surgical world-action modeling have begun to explore motion planning and visual-future prediction within a unified predictive framework. SurgSora and SAW generate future surgical scenes conditioned on prescribed motion cues or tool-tip trajectories, demonstrating the potential of action-guided visual prediction (Chen et al., 2025; Rapuri et al., 2026). Preliminary surgical world-action modeling further predicts future instrument trajectories and visual states jointly from observed surgical history (Huang et al., 2026). However, these advances have not yet established a reliable causal relationship between predicted trajectories and future visual dynamics, which is essential for assessing candidate motions and anticipating risks in robot-assisted interventions. In controllable generation, the trajectory is externally supplied rather than inferred as a motion plan; in joint forecasting, visual states and trajectories may be produced as parallel outputs from the same history without requiring one to influence the other. Consequently, a model can generate a plausible future video and an accurate trajectory while the two predictions remain behaviorally disconnected, as illustrated in Figure 1. Existing evaluations reinforce this

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.

ambiguity by assessing visual quality and trajectory accuracy separately, without testing whether interventions on the predicted trajectory induce spatially corresponding changes in the generated future (Liu et al., 2026).

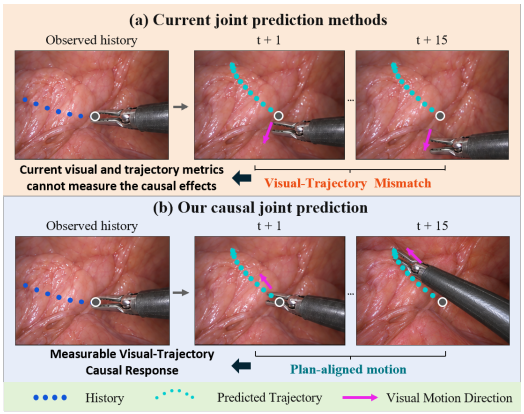

Figure 1: Motivation for trajectory-grounded joint prediction. (a) A joint prediction baseline exhibits a mismatch between its predicted trajectory and visual motion. (b) Our framework generates plan-aligned motion.

Establishing this relationship is particularly important in surgery, where exploratory instrument actions can damage tissue, limiting opportunities to collect intervention data and correct model errors through physical trial and error. This setting raises three key challenges. First, an instrument trajectory is sparse and low-dimensional, whereas surgical images are dominated by complex anatomy and appearance. Small tip displacements can nevertheless alter local tissue contact, deformation, and occlusion. The trajectory condition must therefore be spatially grounded to capture these localized consequences. Second, evaluating candidate motions before execution requires a measurable relationship between trajectory changes and visual responses. Accurate predictions of both outputs alone do not establish this relationship, and standard visual-quality and trajectory-error metrics do not quantify the location or strength of the induced response. Third, the two prediction objectives must be jointly optimized so that visual-future learning improves trajectory prediction and predicted trajectories guide visual generation, while preserving the pretrained video prior.

To address these challenges, we propose a causal visual–trajectory joint prediction framework for surgical motion planning. Given historical surgical frames and instrument positions, a history-only planner first predicts the future instrument trajectory. Instead of treating this trajectory as a global condition or an independent output, we transform it into temporally aligned spatial conditions, bind them to local appearance from the observed scene, and inject them into multiple stages of a pretrained video generator. This establishes an explicit trajectory-to-visual pathway through which the predicted plan guides local future dynamics. A differentiable joint-learning path further allows visual-future prediction to provide feedback to trajectory planning. We additionally introduce a visual–trajectory causal evaluation scheme that estimates instrument motion in generated futures using calibrated ROI dense optical flow. Two metrics, Trajectory-Grounded Motion Alignment (TrajAlign) and Trajectory-Grounded Response Gain (TrajGain), compare this motion with the predicted trajectory to measure whether and to what extent it follows the planned direction.

Experimental results demonstrate the advantages of joint prediction of future surgical videos and instrument trajectories. Spatially grounded trajectory conditioning produces measurable plan-aligned visual motion with minimal visual-quality cost, while differentiable joint learning improves trajectory planning through visual-future prediction.

Our main contributions are summarized as follows:

1. We propose a surgical visual-trajectory joint prediction architecture based on spatially grounded trajectory conditioning, enabling future visual states and instrument trajectories to be predicted within a unified framework while establishing an explicit spatial correspondence between planned trajectories and visual dynamics.

2. We introduce a causal evaluation framework for visual–trajectory relationships, which combines controlled trajectory interventions with localized motion-response measurement to assess whether and to what extent trajectory conditions induce plan-aligned changes in generated visual futures.

3. Experimental results demonstrate that the proposed method achieves promising joint prediction of future surgical videos and instrument trajectories. Joint learning improves trajectory planning, while spatially grounded trajectory conditioning significantly strengthens plan-aligned local motion at only a minimal cost to visual quality.

2 Related Work

2.1 Surgical World Models and Joint Prediction

Surgical world models increasingly connect future visual prediction with instrument motion. SurgSora generates surgical videos from an initial image and user-specified motion cues using object-aware appearance features and optical-flow guidance (Chen et al., 2025). SAW conditions generation on language, scene appearance, tissue affordance masks, and tool-tip trajectories (Rapuri et al., 2026), while Cosmos-H-Surgical-Simulator uses robot kinematics to condition a pretrained surgical video model (Nelson et al., 2026). These methods provide a foundation for generating visual futures under provided motion conditions.

Joint prediction additionally requires inferring future motion alongside visual evolution. VideoVLA jointly predicts actions and future visual outcomes for robotic manipulation (Shen et al., 2026). In surgery, preliminary world-action modeling jointly forecasts visual states and instrument trajectories using a shared encoder and separate prediction heads (Huang et al., 2026). Beyond joint prediction, Rezende et al. (2020) study whether learned models correctly predict the consequences of action interventions, highlighting the importance of causal correctness for model-based planning. In surgical joint prediction, we examine how and to what extent intervening on a predicted trajectory changes local visual dynamics. Our framework connects these predictions through differentiable spatial conditioning, enabling visual feedback to trajectory learning, while paired interventions quantify the direction and strength of the induced visual response.

2.2 Trajectory-Conditioned Video Generation and Evaluation

Trajectory-conditioned video generation studies how spatial motion instructions guide visual synthesis. DragNUWA combines text, image, and trajectory conditions through multiscale motion guidance (Yin et al., 2023). DragAnything associates trajectory controls with object appearance using entity representations and spatial Gaussian maps (Wu et al., 2024b). Tora encodes motion into spatiotemporal representations and injects them into diffusion transformer blocks (Zhang et al., 2025). Building on these ideas, our approach uses response-guided temporal alignment and local observed appearance to connect predicted surgical trajectories with future video features.

Evaluation considers both visual quality and adherence to motion conditions. VBench measures visual and temporal quality (Huang et al., 2024), while trajectory-controlled generation also assesses errors between generated and reference object trajectories (Wu et al., 2024b). SurgWMBench evaluates surgical trajectory prediction and stability under rollout or input perturbations (Liu et al., 2026). We complement these evaluations by measuring local visual motion along the instrument trajectory. TrajAlign and TrajGain characterize directional agreement and relative projected magnitude; paired trajectory interventions with fixed history and generation noise further quantify the direction and strength of the induced visual response.

3 Method

3.1 Task Definition

Let $I^{\mathrm{h}} = (I_1, \ldots, I_{T_{\mathrm{h}}}) \in \mathbb{R}^{T_{\mathrm{h}} \times H \times W \times 3}$ denote the observed RGB frames, and $Q^{\mathrm{h}} = (q_1, \ldots, q_{T_{\mathrm{h}}}) \in \mathbb{R}^{T_{\mathrm{h}} \times 2}$ the corresponding instrument-tip coordinates, with horizontal and vertical positions normalized by image width and height. The superscript h denotes history. Given these observations, we predict a future trajectory $\widehat{P} \in \mathbb{R}^{T_{\mathrm{f}} \times 2}$ and future frames $\widehat{I}^{\mathrm{f}} \in \mathbb{R}^{T_{\mathrm{f}} \times H \times W \times 3}$, where f denotes future. We use $T_{\mathrm{h}} = 5$, $T_{\mathrm{f}} = 15$, and $H \times W = 288 \times 512$. The predictions are coupled through

$$\widehat{P} = f_\psi(Q^{\mathrm{h}}), \qquad \widehat{I}^{\mathrm{f}} = G_\theta(I^{\mathrm{h}}, \widehat{P}; \epsilon), \tag{1}$$

where f_ψ and G_θ are the trajectory predictor and video generator, respectively. The latent noise tensor ϵ has independent standard Gaussian entries and matches the generator's latent-state dimensions. During joint training, video prediction gradients propagate through the trajectory-conditioning pathway to optimize f_ψ alongside trajectory supervision.

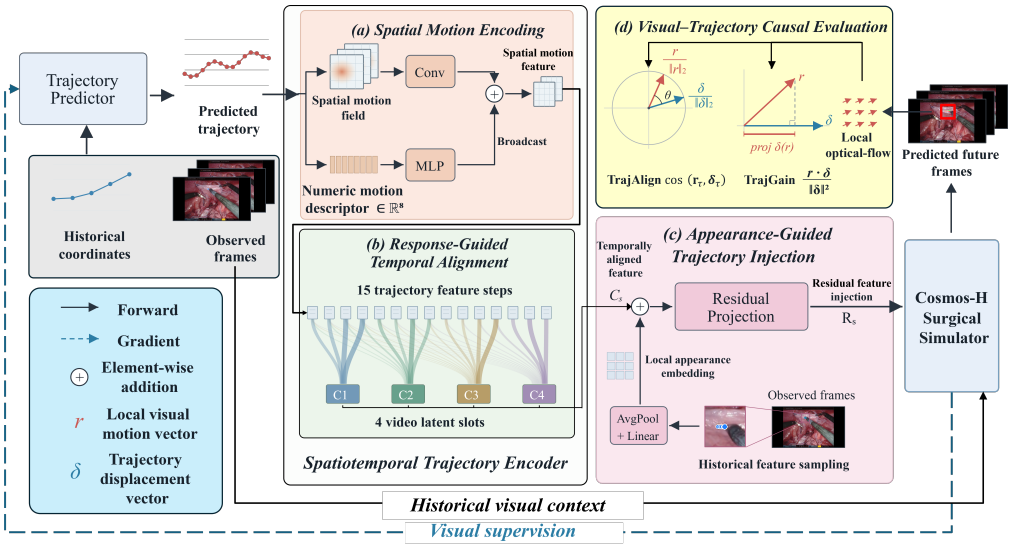

Figure 2: Overview of the proposed framework. Predicted trajectories undergo spatial motion encoding and response-guided temporal alignment, then combine with historical appearance to condition future video prediction. Visual supervision guides trajectory learning, while TrajAlign and TrajGain quantify visual–trajectory alignment and response strength.

3.2 Overview

Built on Cosmos-H-Surgical-Simulator (Nelson et al., 2026), our framework uses instrument-tip trajectories as a spatial and temporal reference for local visual dynamics in surgical scenes. Our working assumption is that visual responses to instrument motion can be assessed near the tip's trajectory, while their appearance depends on the observed instrument and surrounding tissue. This motivates representing predicted trajectories as spatial motion fields and associating them with local historical appearance. Because video compression changes the temporal correspondence, we align these fields with latent slots using measured frame-to-latent responses. Together, these modifications connect predicted motion to corresponding future video features through a differentiable pathway, allowing visual supervision to guide trajectory learning. To assess the resulting trajectory-to-visual influence, we measure local optical flow along the trajectory and use paired interventions to quantify the direction and strength of changes in visual motion.

As illustrated in Figure 2, our framework comprises the following components: a trajectory predictor, a spatiotemporal trajectory encoder, and an appearance-guided trajectory injection. The predictor estimates future instrument-tip positions from historical coordinates. The encoder converts these positions and their displacements into temporally aligned spatial features. The generator combines these features with local appearance from the observed scene to synthesize future frames. During joint training, the trajectory predictor learns from both trajectory accuracy and its contribution to future video prediction, coupling the two tasks through a shared learning objective.

Beyond joint prediction, we aim to quantify the causal influence of trajectory conditioning on future visual dynamics by measuring local motion at corresponding positions and time steps along the instrument trajectory. We introduce TrajAlign and TrajGain to characterize the directional consistency and relative strength of the resulting visual response.

3.3 Trajectory Prediction

The trajectory predictor represents historical instrument motion using both positions and their temporal changes. A Transformer encoder–decoder processes these representations to predict future displacements, which are accumulated from the last observed tip position to reconstruct the future trajectory. This trajectory serves as a differentiable condition for video prediction, enabling visual feedback to complement trajectory supervision during joint training. Appendix A.1 details the input representation and prediction architecture.

3.4 SPATIOTEMPORAL TRAJECTORY ENCODING

A trajectory condition must specify both **where motion should occur** and **how it should evolve over time**. However, combining trajectory points before spatial encoding can merge distinct locations and displacements into an ambiguous condition. We therefore encode each trajectory step separately before aligning the resulting features with the compressed video timeline.

Let $P = (p_1, \ldots, p_{T_\mathrm{f}})$ denote the conditioning trajectory, with $p_0 = q_{T_\mathrm{h}}$ anchoring it to the last observed tip position. For each step, we define the displacement $d_\tau = p_\tau - p_{\tau-1}$ and construct a numerical descriptor

$$a_\tau = [p_\tau,\ d_\tau,\ \|d_\tau\|_2,\ \sin\omega_\tau,\ \cos\omega_\tau,\ m] \in \mathbb{R}^8, \qquad \omega_\tau = \frac{2\pi(\tau-1)}{T_\mathrm{f}}, \tag{2}$$

where the sinusoidal terms encode temporal position and $m \in \{0, 1\}$ indicates whether a trajectory condition is present.

To connect trajectory coordinates with spatial video features, we map each trajectory point onto the backbone's feature grid. We use a Gaussian kernel centered at p_τ to distribute its motion condition over a local neighborhood, providing a smooth spatial representation that remains differentiable with respect to the trajectory position. The resulting map g_τ weights the displacement components and magnitude, forming the spatial motion field:

$$S_\tau = m\,[g_\tau,\ d_\tau^x g_\tau,\ d_\tau^y g_\tau,\ \|d_\tau\|_2 g_\tau,\ m g_\tau] \in \mathbb{R}^{5\times H_\ell \times W_\ell}, \tag{3}$$

where $H_\ell \times W_\ell$ is the feature-grid resolution. A convolutional encoder processes these spatial motion fields, while an MLP encodes the numerical descriptor:

$$E_\tau = \mathrm{Conv}(S_\tau) + \mathrm{Broadcast}(\mathrm{MLP}(a_\tau))\,. \tag{4}$$

The resulting feature E_τ associates each step's position, motion, and temporal identity before temporal aggregation.

Furthermore, the video VAE compresses the temporal dimension, so trajectory steps and video latent slots do not have a one-to-one correspondence. We establish this correspondence through an offline frame-replacement probe of the frozen video encoder using a training clip. Each future frame is individually replaced with its preceding frame, while all other frames remain unchanged. We measure the response $\rho_{s\tau}$ as the mean absolute difference between the original and perturbed encoded features within future latent slot s, averaged over channels and spatial locations. These responses define the support mask $M_{s\tau} = \mathbf{1}[\rho_{s\tau} > 0]$ and initial aggregation weights $\pi_{s\tau} = \rho_{s\tau} / \sum_{j=1}^{T_\mathrm{f}} \rho_{sj}$. The measured supports overlap, allowing each latent slot to incorporate multiple trajectory steps.

We keep M and π fixed and shared across samples. A learnable resampler refines the weights within this measured support, aligning fifteen future trajectory steps with four future latent slots:

$$A_{s\tau} = \frac{M_{s\tau}\pi_{s\tau}\exp(\ell_{s\tau})}{\sum_{j=1}^{T_\mathrm{f}} M_{sj}\pi_{sj}\exp(\ell_{sj})}, \qquad C_s = \sum_{\tau=1}^{T_\mathrm{f}} A_{s\tau} E_\tau, \tag{5}$$

where $\ell_{s\tau}$ are learnable logits initialized to zero, and C_s is the spatial trajectory feature aligned with latent slot s. Thus, the aggregation starts from the measured response weights and adapts during training while preserving their temporal support. Because spatial encoding precedes this aggregation, each trajectory step contributes its encoded location and motion to the corresponding latent slots. The aligned features C_s are subsequently used to condition future video generation.

3.5 APPEARANCE-GUIDED TRAJECTORY INJECTION

The aligned trajectory features describe where and how motion should occur. To connect this geometry with the observed scene, we extract local appearance around the last observed instrument tip. At each selected backbone layer l, we bilinearly sample a 3×3 neighborhood centered at q_{T_h} from the last historical latent slot $H_\mathrm{h}^{(l)}$. The sampled features are averaged and projected into an appearance descriptor:

$$b^{(l)} = W_\mathrm{a}^{(l)}\,\mathrm{AvgPool}\Big(\mathrm{Sample}_{3\times3}\big(H_\mathrm{h}^{(l)}, q_{T_\mathrm{h}}\big)\Big)\,. \tag{6}$$

We broadcast this descriptor across the spatial grid and combine it with each aligned trajectory feature C_s. Two pointwise projections with a SiLU activation produce a residual condition:

$$R_s^{(l)} = W_{\mathrm{o}}^{(l)}\,\mathrm{SiLU}\Big(W_{\mathrm{f}}^{(l)}\left[C_s + \mathrm{Broadcast}\Big(b^{(l)}\Big)\right]\Big), \qquad \widetilde{H}_s^{(l)} = H_s^{(l)} + R_s^{(l)}, \tag{7}$$

where $W_{\mathrm{f}}^{(l)}$ fuses motion and appearance features, and $W_{\mathrm{o}}^{(l)}$ projects them to the backbone's feature dimension. The residual is added only to future latent slots, at four layers distributed across the backbone. We initialize the output projection $W_{\mathrm{o}}^{(l)}$ to zero so that the added pathway initially preserves the pretrained computation. Together with the global trajectory condition, these residual features guide future video generation using both trajectory geometry and local visual context.

3.6 Visual–Trajectory Causal Evaluation

We evaluate the causal influence of trajectory conditioning on future visual dynamics, examining both how generated motion follows the trajectory and how strongly it responds to trajectory conditions. To connect these two modalities, we measure local visual motion at the spatial positions and time steps specified by the instrument trajectory, placing trajectory conditions and generated motion in a common displacement representation.

Let $\bar{p}_\tau$ denote a trajectory point in pixel coordinates and $\delta_\tau = \bar{p}_\tau - \bar{p}_{\tau-1}$ its displacement, with $\bar{p}_0$ corresponding to the last observed instrument tip. We estimate dense optical flow $F_\tau \in \mathbb{R}^{H\times W\times 2}$ between consecutive video frames, using the last observed frame as the starting frame. The local motion response is

$$r_\tau = \frac{1}{|\Omega_\tau|}\sum_{x\in\Omega_\tau} F_\tau(x), \tag{8}$$

where Ω_τ is a fixed-size square region centered at $\bar{p}_\tau$. This associates each trajectory step with a local visual-motion measurement.

We introduce two complementary metrics: TrajAlign measures directional agreement, while TrajGain measures the relative projected response:

$$\mathrm{TrajAlign} = \frac{1}{|\mathcal{M}|}\sum_{\tau\in\mathcal{M}}\frac{r_\tau^\top\delta_\tau}{\|r_\tau\|_2\|\delta_\tau\|_2}, \qquad \mathrm{TrajGain} = \frac{\sum_{\tau\in\mathcal{M}} r_\tau^\top\delta_\tau}{\sum_{\tau\in\mathcal{M}}\|\delta_\tau\|_2^2}. \tag{9}$$

Here, $\mathcal{M}$ contains steps with nonzero trajectory displacement; a zero motion response contributes zero to TrajAlign. TrajAlign approaches one for directionally aligned motion, while TrajGain values below, equal to, or above one indicate attenuated, matched, or amplified projected responses, respectively.

3.7 Joint Training Strategy

We first train the trajectory-conditioning modules using ground-truth trajectories and future videos, with the trajectory predictor and pretrained video backbone fixed. We then use predicted trajectories as conditions and jointly optimize the predictor and conditioning modules through

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{video}} + \lambda_{\mathrm{traj}}\mathcal{L}_{\mathrm{traj}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{video}}$ is the generator's native flow-matching loss and $\mathcal{L}_{\mathrm{traj}}$ supervises trajectory positions, displacements, and the final endpoint. The differentiable conditioning pathway allows visual prediction gradients to update the trajectory predictor alongside direct trajectory supervision. The detailed training procedure and loss definition are provided in Appendix A.2.

4 Experiments

4.1 Experimental Setup

(a) Datasets and Prediction Protocol. We conduct experiments using a multi-source version of SurgWMBench (Liu et al., 2026), comprising annotated surgical clips from SAR-RARP50 (Psychogyios et al., 2023), SurgVU (Zia et al., 2025), and Cholec80 (Twinanda et al., 2016). The dataset

Table 1: Comparison with video-only, trajectory-only, and joint prediction methods on the complete official test set. TrajAlign and TrajGain values closer to 1 indicate better directional alignment and motion amplitude matching, respectively. Ours (w/o visual feedback) retains both training objectives, but prevents the visual prediction loss from updating the trajectory predictor.

Method	Category	Venue	PSNR ↑	SSIM ↑	LPIPS ↓	ADE ↓	FDE ↓	TrajAlign ↑	TrajGain ↑
Cosmos-H Predict (He et al., 2025)	Video-only	arXiv'25	21.0963	**0.7046**	**0.1959**	–	–	–	–
iVideoGPT (Wu et al., 2024a)	Video-only	NeurIPS'24	20.1455	0.6379	0.4408	–	–	–	–
SVD XT1.1 (Blattmann et al., 2023)	Video-only	arXiv'23	20.4820	0.6966	0.2185	–	–	–	–
PECNet (Mangalam et al., 2020)	Trajectory-only	ECCV'20	–	–	–	20.4500	36.9651	–	–
MID (Gu et al., 2022)	Trajectory-only	CVPR'22	–	–	–	27.8813	48.2970	–	–
LED (Mao et al., 2023)	Trajectory-only	CVPR'23	–	–	–	19.1412	34.4310	–	–
VideoVLA (Shen et al., 2026)	Joint	NeurIPS'25	20.6212	0.7020	0.2809	24.7821	44.6860	0.3523	0.1524
Cosmos-Predict2.5 (Ali et al., 2025)	Joint	arXiv'25	20.4383	0.6741	0.2123	27.9646	40.8336	0.0507	0.0066
TSWAM (Huang et al., 2026)	Joint	MICCAI'26	**21.6248**	0.6943	0.5107	18.5450	33.6825	0.0643	-0.0008
V-JEPA2 (Assran et al., 2025)	Joint	arXiv'25	21.2846	0.6849	0.5352	19.6068	35.4211	-0.0311	0.0034
Ours (w/o visual feedback)	Joint	–	20.6975	0.6843	0.2023	18.7564	33.1488	0.4015	0.3625
Ours	Joint	–	20.6691	0.6844	0.2042	**18.5316**	**32.4472**	**0.4110**	**0.3934**

covers diverse instrument motions, including needle insertion and extraction during suturing, as well as tissue cutting. It contains 3,044 clips, divided into 2,024 training, 469 validation, and 551 test clips, with the test set covering 56 source videos. Clips from the same source video are kept within the same split.

Each sample consists of 20 frames with corresponding two-dimensional instrument-tip coordinates. We use the first five frames and their coordinates as observations and predict the subsequent fifteen frames and tip positions. Images are resized to 288×512, and coordinates are normalized by image width and height. At inference, the model receives only the observed frames and historical coordinates; future frames and trajectories are used exclusively for evaluation.

(b) Evaluation Metrics. Following SurgWMBench (Liu et al., 2026), we evaluate visual prediction using PSNR, SSIM, and LPIPS, and trajectory prediction using ADE and FDE. We additionally use TrajAlign and TrajGain as defined in Eq. 9 to measure directional consistency and relative response strength, both for trajectory following and paired causal responses. Detailed evaluation procedures and settings are provided in Appendix A.3.

(c) Implementation Details. We train each of the two stages for 20 epochs using AdamW with a batch size of 8, setting $\lambda_{\mathrm{traj}} = 0.05$ during joint training. Detailed optimization, architecture, and optical-flow settings are provided in Appendix A.4.

4.2 Comparison with State-of-the-Art

As shown in Table 1, we compare our framework with video-only, trajectory-only, and joint prediction methods to assess individual prediction quality and visual–trajectory consistency. We also include a variant without visual feedback to isolate its contribution to trajectory learning.

Against single-task baselines, our framework achieves higher PSNR and lower LPIPS than iVideoGPT and SVD, although Cosmos-H Predict remains stronger across all three visual metrics. Meanwhile, our method outperforms all trajectory-only baselines, reducing ADE and FDE by 3.18% and 5.76%, respectively, relative to LED.

Among existing joint prediction baselines, our framework achieves the lowest ADE, FDE, and LPIPS, together with the highest TrajAlign (0.4110) and the TrajGain closest to unity (0.3934). Its ADE is nearly tied with TSWAM, while its FDE is 3.67% lower. TSWAM and V-JEPA2 achieve higher PSNR but substantially lower motion-consistency scores, showing that visual reconstruction quality alone does not ensure agreement with predicted trajectories.

The variant without visual feedback retains the same architecture and both training objectives, but prevents the visual loss from updating the trajectory predictor. Enabling this feedback reduces ADE and FDE by 1.20% and 2.12%, respectively, improves TrajAlign, and brings TrajGain closer to unity, while slightly worsening PSNR and LPIPS. This controlled comparison supports a modest benefit of visual supervision for trajectory prediction and visual–trajectory consistency.

Table 2: Component ablation on official test set. All settings use joint training. S: Spatial Motion Encoding; A: Appearance-Guided Trajectory Injection; T: Response-Guided Temporal Alignment. Disabling T uses ordinary contiguous temporal grouping with learnable within-group weights.

S	A	T	PSNR ↑	SSIM ↑	LPIPS ↓	ADE ↓	FDE ↓	TrajAlign ↑	TrajGain ↑
–	–	–	20.3806	0.6742	0.2101	18.6104	32.6013	0.3259	0.3199
✓	–	–	20.4556	0.6776	0.2077	19.5238	34.2006	0.3354	0.3624
–	✓	–	20.5900	0.6807	0.2050	18.6712	32.7070	0.3700	0.3721
–	–	✓	20.3761	0.6739	0.2102	18.5983	32.5868	0.3271	0.3168
✓	✓	–	**20.7257**	**0.6855**	**0.2015**	18.6593	32.7126	0.3888	0.3764
✓	–	✓	20.4328	0.6758	0.2085	18.6400	32.6783	0.3591	0.3290
–	✓	✓	20.6617	0.6841	0.2030	18.7298	33.0992	0.3865	0.3390
✓	✓	✓	20.6691	0.6844	0.2042	**18.5316**	**32.4472**	**0.4110**	**0.3934**

4.3 Ablation Study

Table 2 compares all eight combinations of Spatial Motion Encoding (S), Appearance-Guided Trajectory Injection (A), and Response-Guided Temporal Alignment (T) under joint training. Adding A alone improves visual quality, TrajAlign, and TrajGain, with a small increase in trajectory errors. Adding S alone also improves visual and motion-consistency metrics but increases ADE and FDE. Using T alone yields little change relative to ordinary contiguous temporal grouping with learnable within-group weights.

Combining S and A achieves the best visual metrics. Further incorporating T increases TrajAlign from 0.3888 to 0.4110 and TrajGain from 0.3764 to 0.3934, corresponding to relative improvements of 5.70% and 4.53%. ADE and FDE decrease by 0.68% and 0.81%, respectively, alongside a small decline in visual metrics. These results suggest that response-guided temporal alignment is most beneficial when combined with spatial motion encoding and appearance-guided trajectory injection. The full model achieves the lowest trajectory errors and the strongest visual–trajectory consistency among the evaluated configurations.

4.4 Visual–Trajectory Causal Response Analysis

We examine whether trajectory conditioning has a measurable and directionally consistent causal influence on future visual dynamics. We conduct paired trajectory interventions on 64 randomly sampled non-training clips. For each clip, we independently apply four rotations ($\pm 30^\circ$ and $\pm 60^\circ$) and four displacement scales (0, 0.5, 1.5, and 2), anchored at the last observed instrument-tip position. Each intervention is applied to the original trajectory and compared with the unmodified condition; rotations and scaling are tested separately. All eight comparisons are repeated with three sampling seeds per clip. Within each pair, only the trajectory supplied to the spatial conditioning pathway changes; the observed history, initial noise, model parameters, and global trajectory condition remain fixed.

At each future step, we measure both generated videos over the same pixels, using the union of local regions around the original and intervened trajectory positions. This captures motion near both positions without changing the measurement region between outputs. Let $r_\tau^{(a)}$ and $r_\tau^{(b)}$ denote the mean optical-flow vectors in this shared region, and let $\delta_\tau^{(a)}$ and $\delta_\tau^{(b)}$ denote the corresponding consecutive-step trajectory displacements. We compute $\Delta r_\tau = r_\tau^{(b)} - r_\tau^{(a)}$ and $\Delta\delta_\tau = \delta_\tau^{(b)} - \delta_\tau^{(a)}$, representing the induced change in local visual motion and the requested change in trajectory motion, respectively. Substituting these differences into the TrajAlign and TrajGain formulas yields $\mathrm{TrajAlign}_\Delta$ and $\mathrm{TrajGain}_\Delta$, which quantify the directional consistency and relative projected magnitude of the induced response.

To quantify uncertainty, we resample source-video groups with replacement and recompute the mean response 10,000 times. The 2.5th and 97.5th percentiles give the 95% confidence interval, while grouping preserves dependence between clips from the same video.

As shown in Table 3, $\mathrm{TrajAlign}_\Delta$ reaches 0.4600, with its confidence interval entirely above zero, indicating that induced visual motion changes generally follow the requested trajectory changes. $\mathrm{TrajGain}_\Delta = 0.3717$ indicates an average projected response of approximately 37% of the trajec-

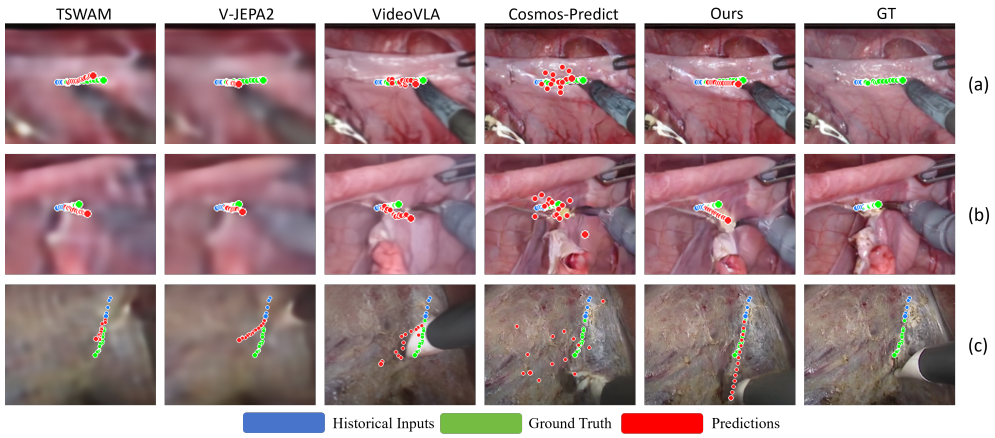

Figure 3: Qualitative comparison under (a) normal joint prediction, (b) 30° trajectory rotation, and (c) 2× displacement scaling. Blue, green, and red points indicate historical positions, factual future positions, and predicted or intervened trajectories, respectively. Ground-truth frames provide factual references.

tory displacement change, revealing a positive but attenuated response. These paired measurements provide evidence that spatial trajectory conditioning causally influences future visual dynamics and quantify the direction and strength of this influence. Complementary conditioning ablations in Appendix A.5 examine how removing or temporally shuffling trajectory conditions affects image fidelity and motion consistency.

4.5 Qualitative Analysis

Table 3: Paired causal responses. Means and 95% confidence intervals are computed using 10,000 bootstrap resamples grouped by source video.

Metric	Interventions	Mean	95% CI
$\text{TrajAlign}_{\Delta}$	Rotations	0.4600	[0.4338, 0.4853]
TrajGain_{Δ}	Rotations and scales	0.3717	[0.3209, 0.4173]

Figure 3 compares joint predictions under three trajectory settings. Under normal prediction in Figure 3(a), our method preserves instrument structure and tissue appearance while producing trajectories that closely follow the ground truth. The baseline outputs exhibit varying degrees of visual blur or trajectory deviation. Our predictions combine visual fidelity with a coherent correspondence between instrument position and predicted trajectory.

Figure 3(b) and (c) examine responses to a 30° trajectory rotation and 2× displacement scaling, respectively. In the rotation example, the generated instrument position is consistent with the redirected trajectory. In the scaling example, the instrument position follows the extended trajectory toward its endpoint. Several baseline outputs remain less consistent with their modified trajectory conditions. Together, these examples illustrate how trajectory conditioning shapes the direction and spatial extent of future instrument motion, complementing the quantitative response analysis.

5 Conclusion

We presented a joint visual–trajectory prediction framework for surgical motion planning that connects predicted instrument trajectories to future visual dynamics through spatial motion encoding, response-guided temporal alignment, and appearance-guided injection. Joint training enables visual supervision to improve trajectory learning, yielding better trajectory accuracy and visual–trajectory consistency while maintaining competitive visual quality. Through TrajAlign, TrajGain, and controlled paired interventions, we quantify the directional consistency and relative strength of visual responses to trajectory changes. These results provide evidence that trajectory conditioning causally influences local future visual dynamics, supporting the study of instrument motion and its predicted visual consequences. Together, the framework and evaluation protocol provide a foundation for studying candidate instrument trajectories and their associated visual consequences in surgical motion planning.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript's language. The authors take responsibility for the final content.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

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

Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17092–17101. IEEE, 2022.

Yufan He, Pengfei Guo, Mengya Xu, Zhaoshuo Li, Andriy Myronenko, Dillan Imans, Bingjie Liu, Dongren Yang, Mingxue Gu, Yongnan Ji, et al. Cosmos-h-surgical: Learning surgical robot policies from videos via world modeling. *arXiv preprint arXiv:2512.23162*, 2025.

Weiliang Huang, Huanrong Liu, Bob Zhang, Qi Dou, Zhen Chen, Yun Gu, Guy Rosman, and Qingbiao Li. Towards surgical world-action modeling: A preliminary joint visual-trajectory forecasting for surgical motion planning. *arXiv preprint arXiv:2608.20284*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

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.

Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision*, pp. 759–776. Springer, 2020.

Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5517–5526. IEEE, 2023.

Nigel Nelson, Juo-Tung Chen, Jesse Haworth, Xinhao Chen, Lukas Zbinden, Dianye Huang, Alaa Eldin Abdelaal, Alberto Arezzo, Ayberk Acar, Farshid Alambeigi, et al. Open-h-embodiment: A large-scale dataset for enabling foundation models in medical robotics. *arXiv preprint arXiv:2604.21017*, 2026.

Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.

Sampath Rapuri, Lalithkumar Seenivasan, Dominik Schneider, Roger Soberanis-Mukul, Yufan He, Hao Ding, Jiru Xu, Chenhao Yu, Chenyan Jing, Pengfei Guo, et al. Saw: Toward a surgical action world model via controllable and scalable video generation. *arXiv preprint arXiv:2603.13024*, 2026.

Danilo J Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, et al. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836*, 2020.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. Videovla: Video generators can be generalizable robot manipulators. *Advances in neural information processing systems*, 38:95597–95621, 2026.

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, 2016.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024a.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024b.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2063–2073. IEEE, 2025.

Aneeq Zia, Max Berniker, Rogerio Nespolo, Xiaorui Zhang, Conor Perreault, Ziheng Wang, Benjamin Mueller, Ryan Schmidt, Kiran Bhattacharyya, Xi Liu, et al. Surgical visual understanding (surgvu) dataset. *arXiv preprint arXiv:2501.09209*, 2025.

A Appendix

A.1 Trajectory Predictor

The trajectory predictor takes historical instrument-tip coordinates $Q^{\mathrm{h}} = (q_1, \ldots, q_{T_{\mathrm{h}}})$ as input, where $q_t \in \mathbb{R}^2$ contains horizontal and vertical positions normalized by image width and height, respectively. It predicts T_{f} future tip positions using only these historical coordinates.

Historical motion representation. We represent each observation using its absolute position, position relative to the last observation, displacement, and displacement change:

$$u_t = [q_t,\ q_t - q_{T_{\mathrm{h}}},\ v_t,\ a_t] \in \mathbb{R}^8, \tag{11}$$

where $v_t = q_t - q_{t-1}$ for $t \geq 2$ and $a_t = v_t - v_{t-1}$ for $t \geq 3$. We set $v_1 = 0$ and $a_1 = a_2 = 0$. An MLP projects these descriptors into motion tokens. The tokens are combined with learnable temporal embeddings and processed by a Transformer encoder.

Future trajectory prediction. A Transformer decoder uses learnable future queries to attend to the encoded history and produce future motion features z_τ. A prediction head g_ψ maps each feature to a bounded displacement:

$$\hat{d}_\tau = \delta_{\max} \tanh(g_\psi(z_\tau)), \qquad \hat{p}_\tau = q_{T_{\mathrm{h}}} + \sum_{j=1}^{\tau} \hat{d}_j, \qquad \tau = 1, \ldots, T_{\mathrm{f}}, \tag{12}$$

where $\delta_{\max}$ bounds each displacement component in normalized image coordinates. Accumulating the predicted displacements yields the future trajectory $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_{T_{\mathrm{f}}})$. This trajectory is subsequently encoded into video conditions. During joint training, gradients from both the trajectory objective and the video prediction objective update the predictor through this differentiable computation.

A.2 Joint Training Details

We train the framework in two stages: trajectory-conditioning training and joint visual–trajectory training.

Trajectory-conditioning training. In the first stage, we train the spatial trajectory-conditioning modules using paired ground-truth trajectories and future videos. The trajectory predictor and the inherited video backbone remain fixed. Ground-truth trajectories are processed by the spatiotemporal trajectory encoder and appearance-guided injection modules to condition future video prediction. These modules are optimized using the generator's native flow-matching objective $\mathcal{L}_{\mathrm{video}}$, establishing the correspondence between trajectory conditions and future visual dynamics.

Joint visual–trajectory training. In the second stage, we replace the ground-truth trajectory condition with the predicted trajectory $\hat{P} = f_\psi(Q^{\mathrm{h}})$. We jointly optimize the trajectory predictor and trajectory-conditioning modules while keeping the inherited video backbone fixed. The joint objective in Equation 10 combines $\mathcal{L}_{\mathrm{video}}$ with the trajectory loss

$$\mathcal{L}_{\mathrm{traj}} = \mathrm{MAE}(\hat{P}, P^*) + \frac{1}{2}\,\mathrm{MAE}(\Delta\hat{P}, \Delta P^*) + \frac{1}{2}\,\mathrm{MAE}(\hat{p}_{T_{\mathrm{f}}}, p^*_{T_{\mathrm{f}}}), \tag{13}$$

where P^* is the ground-truth future trajectory, ΔP contains differences between consecutive future positions, and MAE averages absolute errors over all entries. The three terms supervise future positions, inter-step displacements, and the final endpoint, respectively. The coefficient λ_{traj} balances trajectory and video supervision.

Visual feedback to trajectory learning. The predicted trajectory remains differentiable throughout spatial motion encoding, temporal alignment, and appearance-guided injection. Consequently, gradients from $\mathcal{L}_{\mathrm{video}}$ propagate through the video generator and conditioning modules to the trajectory predictor. Its parameters are therefore updated by both direct trajectory supervision and the effect of its predictions on future video prediction.

A.3 Evaluation Details

Video and trajectory prediction. PSNR, SSIM, and LPIPS are computed over the fifteen predicted frames at a resolution of 288×512 and averaged across frames and test samples. Average Displacement Error (ADE) measures the mean Euclidean distance between predicted and ground-truth instrument-tip positions over the prediction horizon. Final Displacement Error (FDE) measures this distance at the final prediction step. Both trajectory errors are reported in pixels at the evaluation resolution.

Visual–trajectory evaluation. We apply the TrajAlign and TrajGain definitions from the methodology in two settings. For trajectory following, we compare local visual motion in generated videos with the corresponding trajectory displacements. For paired causal responses, we instead compare the changes in local visual motion with the changes in trajectory displacement under controlled interventions, denoting these scores as $\mathrm{TrajAlign}_{\Delta}$ and $\mathrm{TrajGain}_{\Delta}$. Higher alignment scores indicate stronger directional agreement. Gain values closer to one indicate better matching of the projected response magnitude to the corresponding trajectory displacement; negative values indicate an opposing response, values between zero and one indicate attenuation, and values above one indicate amplification.

A.4 Implementation Details

Optimization. The trajectory-conditioning stage and the joint prediction stage are each trained for 20 epochs using AdamW with a batch size of 8. The learning rate is 10^{-4} for the trajectory-conditioning modules and 5×10^{-6} for the trajectory predictor during joint training. We use linear warm-up over the first 8% of optimization steps, followed by cosine decay to 10^{-6}. Weight decay is set to zero, and the gradient norm is clipped to 1.0. The trajectory loss weight is fixed at $\lambda_{\mathrm{traj}} = 0.05$ following initial gradient-norm calibration. Trajectory conditions are dropped with probability 0.1 during training. Each run uses a single NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of memory.

Conditioning architecture. The spatial conditioning features have 128 channels. Gaussian kernels use a standard deviation of 24 pixels at the 288×512 reference resolution. When mapped to feature grids, the kernel scale is adjusted separately along the horizontal and vertical axes. Local appearance features are sampled from a 3×3 neighborhood around the last observed instrument-tip position. Residual conditions are injected into future latent slots at four layers distributed across the video backbone.

Optical-flow settings. For visual–trajectory evaluation, we estimate dense optical flow using the Farnebäck method. Local motion is measured within square regions with a half-width of 24 pixels at the 288×512 evaluation resolution.

A.5 Trajectory Conditioning Ablation

We compare predicted, missing, and temporally shuffled trajectory conditions to examine their effects on image fidelity and visual–trajectory consistency. The observed history and initial generation noise are fixed across conditions. Missing conditioning removes the predicted future trajectory from the video conditions, while shuffling temporally permutes the predicted positions. The predicted and missing conditions are evaluated using the original predicted trajectory and its local measurement regions; the shuffled condition is evaluated using the shuffled trajectory and its corresponding regions.

As shown in Table 4, providing the predicted trajectory improves all three visual-quality metrics over missing conditioning and increases TrajAlign and TrajGain by 0.1126 and 0.0922, respectively. Temporally shuffled conditioning retains similar image fidelity but substantially reduces agreement with the supplied trajectory: TrajAlign falls to 0.0435 and TrajGain to 0.0098. These results show that image fidelity and trajectory consistency capture distinct properties. They also demonstrate that the predicted trajectory and its temporal order contribute to generating visual motion consistent with the plan.

Table 4: Trajectory-conditioning ablation with fixed history and initial generation noise. Missing conditioning removes the predicted future trajectory from the video conditions; shuffling temporally permutes its positions. TrajGain values closer to one indicate better projected motion-amplitude matching.

Trajectory condition	PSNR ↑	SSIM ↑	LPIPS ↓	TrajAlign ↑	TrajGain ↑
Predicted plan	**20.6691**	**0.6844**	**0.2042**	**0.4110**	**0.3934**
Missing condition	20.4484	0.6783	0.2092	0.2984	0.3012
Shuffled plan	20.6555	0.6843	0.2045	0.0435	0.0098

A.6 Baseline Adaptations

We adapt the comparison methods to the five-observation, fifteen-step surgical prediction task. Video-only methods receive historical RGB frames, trajectory-only methods receive historical tip coordinates, and joint methods receive both. All methods are evaluated on the same 551 test clips. Table 5 summarizes their task-specific adaptations and trainable components.

Table 5: Task adaptations of the comparison baselines. The table describes the implementations evaluated in this study.

Method	Task adaptation	Training configuration
Video-only methods		
Cosmos-H Predict	The released surgical video predictor is conditioned on five historical frames, without a trajectory head or trajectory loss.	Attention LoRA with rank 16; frozen pretrained base and VAE; native flow-matching objective.
iVideoGPT	The action-free pretrained model is adapted to five-frame context and fifteen-frame autoregressive prediction.	Pretrained Transformer fine-tuning with token cross-entropy; frozen tokenizer.
SVD XT1.1	The input convolution is expanded to accept five historical latent maps. Historical CLIP embeddings are averaged to provide image conditioning.	Expanded input convolution and rank-16 attention LoRA; frozen VAE and CLIP encoder; native EDM objective.
Trajectory-only methods		
PECNet	Observation and prediction lengths are set to five and fifteen. Endpoint-conditioned prediction is retained, with a self-only social mask for independently tracked instruments.	The adapted model is trained from scratch using endpoint, trajectory, and latent regularization losses.
MID	A single-instrument history encoder conditions trajectory diffusion. Predicted displacements are integrated into future positions, without cross-sample social interactions.	The history encoder and diffusion model are trained from scratch.
LED	The denoiser and initializer are adapted to five observed and fifteen future positions, with independent instrument tracks. A predeclared candidate is used at evaluation.	The denoiser is trained first and then frozen during initializer training.
Joint prediction methods		
VideoVLA	Robot actions are replaced by fifteen two-dimensional offsets from the last observed tip position. Task-specific action projections and historical-coordinate conditioning are added while retaining joint video–action denoising.	Rank-16 attention/MLP LoRA and task adapters; frozen pretrained base, VAE, and text encoder.

Table 5 (continued)

Method	Task adaptation	Training configuration
Cosmos-Predict2.5	An MLP reads pooled history-only backbone features and historical coordinates to predict future tip positions. Video and trajectory prediction share trainable backbone parameters; the trajectory output is not fed into video generation.	Rank-16 attention LoRA and a new trajectory head; video flow-matching and trajectory MSE objectives.
TSWAM	Frozen SurgMotion ViT-L features feed a temporal–spatial joint prediction head and an RGB decoder. Three-step predictions are rolled out five times.	An RGB decoder is trained first and frozen; the joint prediction head is then trained.
V-JEPA2	Frozen V-JEPA2 ViT-L features use the same joint-head and RGB-decoder architectures as the TSWAM adaptation, with the same three-step autoregressive rollout.	A separate RGB decoder and joint head are trained using the same two-stage procedure as TSWAM.

Resolution and conditioning. iVideoGPT, VideoVLA, TSWAM, and V-JEPA2 produce 256×256 frames, which are resized to the common 288×512 evaluation resolution. Cosmos-H Predict and Cosmos-Predict2.5 use fixed zero text contexts, while VideoVLA uses the constant prompt "A surgical endoscopic video." Inference uses historical observations only.

Training and model selection. Each trainable stage uses a twenty-epoch budget. LED trains its denoiser and initializer in separate stages; TSWAM and V-JEPA2 separately train their RGB decoders before joint prediction training. Checkpoints are selected using a fixed set of 64 validation clips: maximum PSNR for video-only methods, minimum ADE for trajectory-only methods, minimum joint validation loss for VideoVLA and Cosmos-Predict2.5, and $\mathrm{MSE}_{\mathrm{pixel}} + 10^{-6}\mathrm{ADE}_{\mathrm{px}}$ for TSWAM and V-JEPA2. Test evaluation uses one prediction per clip without ground-truth-based candidate selection.

A.7 Component Ablation Definitions

The eight configurations in Table 2 vary three operations: spatial motion-field encoding (S), local appearance conditioning within trajectory injection (A), and response-guided temporal alignment (T). Numerical trajectory encoding, the inherited global trajectory condition, and the joint training objective are retained in all configurations. Table 6 specifies the corresponding switches.

Spatial and appearance ablations. The S switch controls the spatial motion-field contribution within the trajectory encoder. When S is disabled, the per-step feature becomes $E_\tau = \mathrm{Conv}(0) +$

Table 6: Definitions of the component switches. Disabling a component changes its information contribution or temporal correspondence while retaining the shared prediction and residual-injection framework.

Component	Enabled	Disabled
S	Encode the five-channel spatial motion field constructed from trajectory positions and displacements.	Set the spatial motion field S_τ to zero before convolutional encoding. Retain the convolutional module and numerical trajectory encoding.
A	Extract the local historical appearance descriptor $b^{(l)}$ and combine it with aligned trajectory features.	Set $b^{(l)}$ to zero. Retain the aligned trajectory features, residual projection, and injection into future video features.
T	Use the measured temporal support and response-based priors, with learnable weights within the support.	Use contiguous groups 1–4, 5–8, 9–12, and 13–15, with uniform initial priors and learnable within-group weights.

$\mathrm{Broadcast}(\mathrm{MLP}(a_\tau))$. The convolutional layers, including their biases, are retained. When A is disabled, the local appearance descriptor is set to zero after projection, while the residual injection continues to operate on the aligned trajectory features.

Ordinary temporal grouping. When T is disabled, we replace the measured support with four non-overlapping groups, $\mathcal{G}_1 = \{1, \ldots, 4\}$, $\mathcal{G}_2 = \{5, \ldots, 8\}$, $\mathcal{G}_3 = \{9, \ldots, 12\}$, and $\mathcal{G}_4 = \{13, \ldots, 15\}$. We set $M_{s\tau} = \mathbf{1}[\tau \in \mathcal{G}_s]$ and $\pi_{s\tau} = 1/|\mathcal{G}_s|$ within each group. The learnable logits are initialized to zero and optimized using the same aggregation formula as the full model.

Configuration with all three components disabled. This configuration retains the trajectory predictor, numerical trajectory encoding, ordinary temporal grouping, residual injection, and inherited global trajectory conditioning. Both trajectory supervision and visual feedback remain active. It therefore provides the reference configuration for measuring the additional contributions of S, A, and T.