

SurgCast: Action-Conditioned Future Skeletons for Controllable Surgical World Models

Anonymous Authors

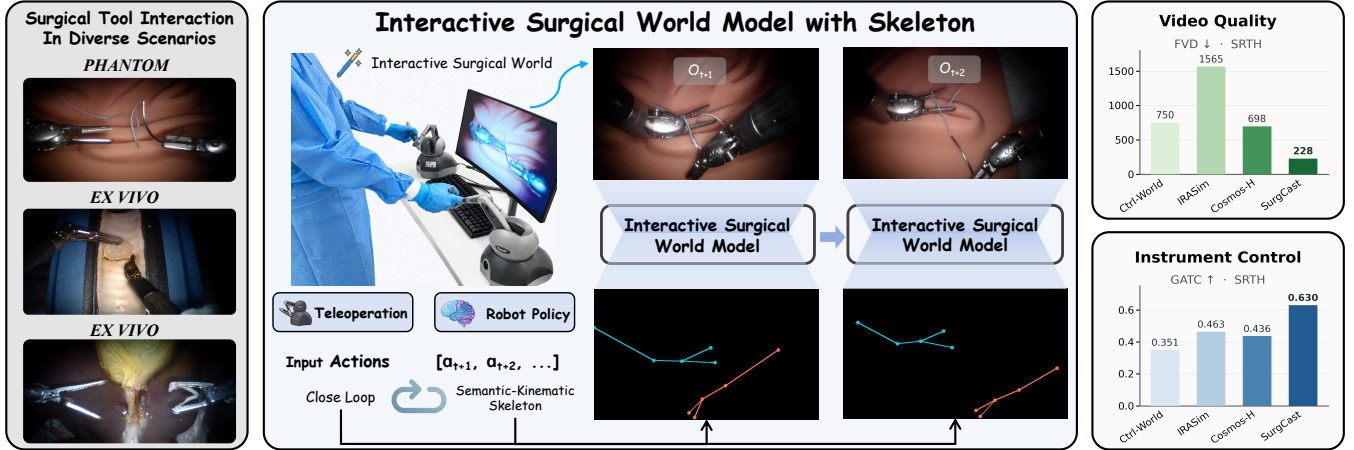


Fig. 1. SurgCast forecasts future instrument skeletons from prospective robot actions and uses their geometry and derived kinematics to guide video generation. Read from surgical scenes (left), through commands and their structural and visual predictions (center), to quality and control summaries (right). The illustrated robot-policy loop motivates the framework; the evaluations use supplied action windows.

Abstract—Surgical world models must turn prospective robot actions into visually faithful, structurally consistent instrument motion. We introduce SurgCast, which predicts future instrument skeletons from current structure, robot state, and future actions, then uses the forecast to control video generation. Dual-path injection combines skeleton geometry with derived, instrument-specific motion descriptors through geometric residuals and conditioning modulation. Both paths complement native action conditioning and are retained in a four-step causal student. On SutureBot and SRTH-Porcine-Cholei, the teacher and student improve all seven reported quality and instrument-consistency metrics over their respective direct-action baselines; teacher FVD decreases by 37.6% and 67.3%. Ablations show that future actions improve skeleton prediction and that combining both injection paths performs best across all six metrics under predicted skeletons. These results support using predicted articulated structure and its motion semantics as a control interface for surgical video prediction. Project videos and examples are available at <https://surgecast.netlify.app/>.

I. INTRODUCTION

A surgical robot’s actions determine how instruments move within the operative scene. An action-conditioned world model should make these consequences visible before execution: different commands from the same initial view should produce different, structurally consistent motions. Visual prediction supports robot planning and interactive simulation [1], [2], [3], while action-conditioned models extend video generation to manipulation and surgery [4], [5], [6]. For surgery, realistic appearance must accompany instrument configurations that follow the supplied commands.

Direct action conditioning leaves the generator to infer future articulated geometry in image space. Our qualita-

tive comparisons show that recognizable scenes can coexist with instrument drift and interaction misalignment (Fig. 3). Trajectory and kinematic controls expose structure to video models [7], [8], [9], [10], [11]. Building on this principle, we study an explicit predictive interface: forecast the instrument configuration from current structure and prospective commands, then use that forecast to guide video generation.

Instrument skeletons connect these stages through keypoints with fixed tool-specific roles. Their trajectories also yield descriptors of image-plane translation, shaft-direction angular velocity, and projected wrist and jaw articulation. We call these instrument-meaningful motion descriptors *kinematic semantics*. Geometric maps locate the articulated tool, while derived descriptors expose its motion directly to the generator’s conditioning. This raises two questions: do future commands improve skeleton prediction, and does using the forecast’s geometry and kinematic semantics improve action-conditioned video generation?

SurgCast addresses both questions. An Action-to-Future-Skeleton Predictor forecasts keypoints and visibility from current structure, robot state, and future actions. Dual-Path Semantic-Kinematic Injection converts the same forecast into geometry maps for spatial and temporal residuals and motion descriptors for conditioning modulation. The generator retains its native action inputs. This interface conditions a full-sequence teacher and a four-step causal student trained through distribution matching [12], [13]. Our contributions are:

- An action-to-structure-to-video framework that uses predicted future instrument skeletons as an explicit

control interface (Fig. 2; Table III).

- Dual-path injection of skeleton geometry and its derived motion semantics, with complementary benefits demonstrated under predicted skeletons (Table II).
- Retention of the interface in a four-step causal student, with improved video quality and instrument consistency over the corresponding direct-action baselines (Table I; Fig. 5).

II. RELATED WORK

A. Action-Conditioned Prediction for Robotics

Visual foresight, generative policies, and interactive simulators connect visual prediction to robot planning [1], [2], [3]. iVideoGPT and DINO-WM study interactive prediction and planning in visual feature spaces [14], [15]; IRASim, HMA, and Ctrl-World model action-conditioned visual dynamics [4], [16], [5]. In surgery, Open-H-Embodiment supplies paired video and kinematics [17], and Cosmos-H-Dreams provides action-conditioned generation and causal deployment [6]. Instrument pose and tracking supply articulated image-space structure [18], [19], [20], while forecasting methods predict motion or state from historical observations [21], [22]. SurgCast predicts future structure conditioned on prospective commands and uses it as an intermediate control for video generation.

B. Structured Control for Surgical Video Generation

ControlNet and T2I-Adapter introduce spatial diffusion conditioning [23], [24]; VideoComposer, DragNUWA, and MotionCtrl extend control to video motion [7], [8], [9]. Surgical generators use action graphs, latent actions, object-aware features, and tool-tip trajectories [25], [26], [27], [10]. KVLR projects articulated kinematics using camera parameters into image-aligned semantic, depth, orientation, velocity, and acceleration controls and routes these modalities hierarchically [11]. Our control construction first forecasts future image-space skeletons from current structure and commands. The forecast then supplies both geometric maps and instrument-motion descriptors, injected as residuals and conditioning modulation, respectively.

C. Few-Step Causal Video Diffusion

Progressive and distribution-matching distillation reduce sampling cost [28], [29], [12]. Diffusion Forcing and CausVid study sequence and causal generation [30], [13]; Self Forcing, Rolling Forcing, and Causal Forcing address self-generated histories and extended rollouts [31], [32], [33]. SurgCast uses this training framework to retain its predicted structural interface in a four-step causal generator.

III. METHOD

A. Problem Formulation and Overview

Given an initial image I_0 , current instrument skeleton S_0 , robot state r_0 , and prospective action sequence $a_{1:H}$, SurgCast generates future observations $I_{1:H}$ whose instrument configuration follows the supplied commands. Each instrument is represented by five image-space keypoints with

fixed structural roles and visibility indicators. We denote the two-instrument structure by $S_h = (P_h, V_h)$, with $P_h \in \mathbb{R}^{2 \times 5 \times 2}$. The prediction and generation stages are

$$\begin{cases} \hat{S}_{1:H} = F_\phi(S_0, r_0, a_{1:H}), \\ I_{1:H} \sim p_\theta(\cdot \mid I_0, a_{1:H}, \mathcal{C}(S_0, \hat{S}_{1:H})). \end{cases} \quad (1)$$

The operator $\mathcal{C}$ constructs two controls from the same forecast: rasterized skeleton geometry and interpretable kinematic descriptors derived from its keypoint trajectories (Fig. 2). These descriptors express image-plane translation, shaft-direction angular velocity, and projected wrist and jaw articulation. The same structural interface conditions the full-sequence teacher and causal student, both of which retain the backbone’s original action pathway and generate compressed video latents. The current skeleton is a supplied input; confidence, bounding boxes, and tool identities accompany it as predictor inputs.

For long-horizon generation, each sequence contains 73 image frames: one reference and 72 future frames. Temporal compression maps these to 19 video latents, with a separate reference latent and 18 groups of four future frames. Structural conditions are temporally pooled onto this same grid. We use h for image/skeleton time, j for latent time, and σ for diffusion noise level throughout.

B. Action-to-Future-Skeleton Prediction

a) *Anchored structural forecasting*: The predictor uses the current configuration as a common reference for future motion. We normalize keypoint coordinates to $[-1, 1]$. A bounding-box-conditioned completion module first estimates missing current keypoints from visible structure, confidence, robot state, and tool identity. Observed keypoints are preserved; completed points provide anchors for occluded locations with valid bounding boxes. Let $\tilde{P}_0$ denote the resulting current configuration. The predictor embeds this configuration and robot state into a context vector c_0 . At step h , the input token is $u_h = E_a(a_h) + c_0 + e_h$, where e_h is a learned forecast-horizon embedding. A causally masked transformer computes $q_{1:H} = \text{Tr}_\phi(u_{1:H})$. Its output heads predict coordinates and visibility:

$$\begin{cases} \hat{P}_h = \tilde{P}_0 + D_\phi(q_h), \\ \hat{V}_h = \text{sigmoid}(W_v q_h). \end{cases} \quad (2)$$

A causal attention mask makes q_h depend on commands up to step h , with a padding mask for variable-length sequences. Coordinates are predicted as offsets from the completed current configuration instead of accumulated frame-to-frame displacements. The implemented transformer has four layers, width 256, and eight attention heads. Per instrument, the predictor represents commands as three relative translation values, a six-dimensional relative rotation, and one absolute jaw value, normalized using training statistics. The video backbone separately constructs the corresponding 20-dimensional two-arm action representation and pads it to 44 dimensions.

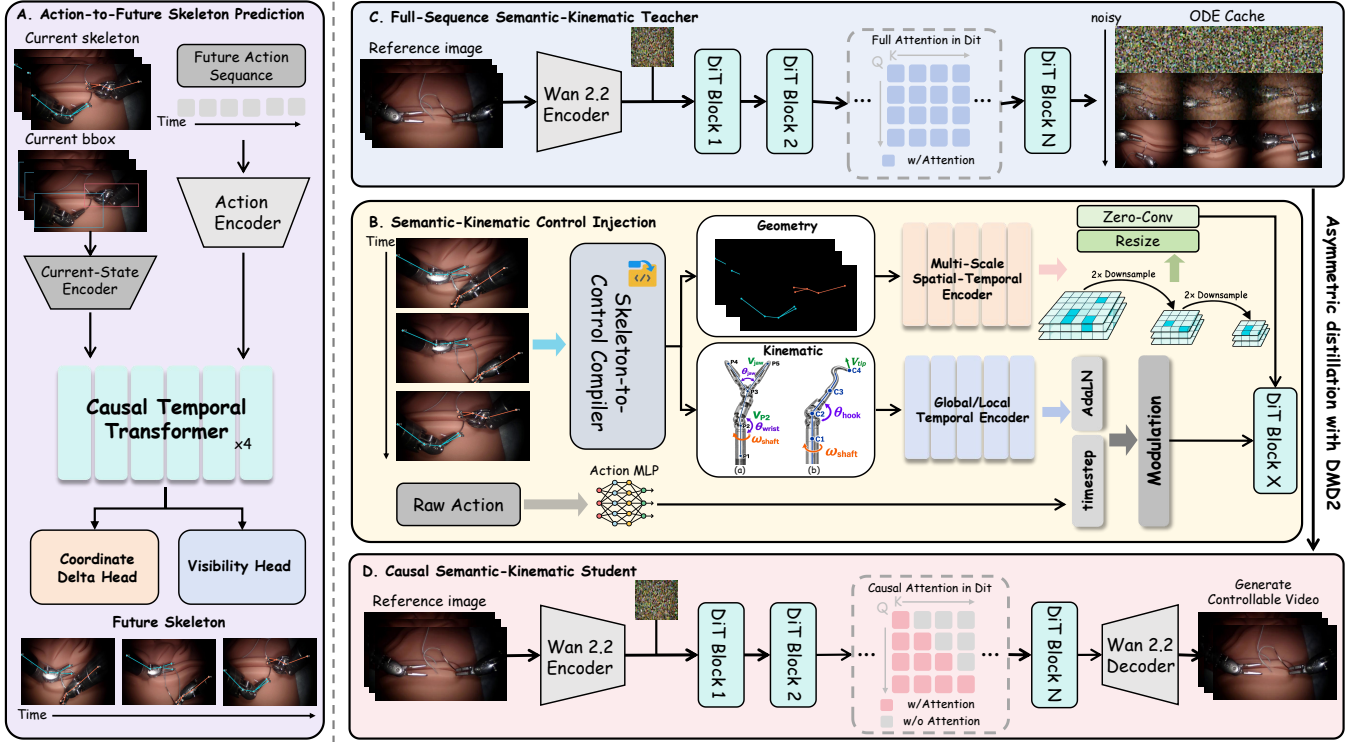


Fig. 2. SurgCast’s action-to-structure-to-video framework. Read (A)→(B) for construction of the control interface: (A) forecasts future instrument skeletons from current structure and prospective actions; (B) converts the forecast into geometry maps for residual injection and derived kinematic descriptors for modulation. Both paths condition the teacher (C) and causal student (D), alongside native action inputs. The arrow from (C) to (D) denotes training-time distillation; teacher and student are alternative video generators. Kinematic quantities describe projected keypoint motion, including shaft-direction angular velocity in the image plane.

b) Visibility-aware learning: The joint prediction objective is

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{coord}} + 0.2\mathcal{L}_{\text{disp}} + 0.1\mathcal{L}_{\text{vis}} + 0.5\mathcal{L}_{\text{comp}}. \quad (3)$$

Coordinate ($\mathcal{L}_{\text{coord}}$) and current-keypoint completion ($\mathcal{L}_{\text{comp}}$) use Smooth- L_1 with threshold 0.02. $\mathcal{L}_{\text{disp}}$ applies the same penalty to consecutive-coordinate differences, using pairs with valid targets. Coordinate supervision is confidence-weighted and normalized within each instrument before combining instrument losses; displacement weights use the minimum confidence of each pair. Visibility loss $\mathcal{L}_{\text{vis}}$ uses binary cross-entropy on known labels. Completion supervision targets artificially masked current keypoints. To handle interchangeable jaw-tip labels, the coordinate loss selects the lower-error identity or jaw-tip-swapped assignment for each instrument over the entire prediction horizon, and applies that assignment consistently to the corresponding targets. The trained predictor supplies structural conditions to video generation.

C. Dual-Path Semantic–Kinematic Injection

The two paths give the generator complementary access to the same predicted skeleton. Geometric maps preserve the spatial support of keypoints and edges and encode their temporal changes as localized residuals. Derived kinematic descriptors expose instrument motion in a compact form for conditioning modulation. This division lets structure guide

video tokens while instrument-meaningful motion quantities guide the conditioning. Global and local descriptors are groups within the kinematic path. Both paths complement the native 44-dimensional action interface.

a) Multi-scale geometric residuals: This path preserves the spatial support and temporal evolution of the skeleton. For each instrument, we rasterize five keypoints and the four edges (1, 2), (2, 3), (3, 4), and (3, 5) as Gaussian point and segment-distance maps. Corresponding validity channels distinguish absent observations from empty spatial support. The two instruments yield 18 geometric channels and 18 validity channels. After temporal alignment, these form $G_j \in \mathbb{R}^{36 \times 36 \times 64}$ on the latent spatial grid.

Separate convolutional stems encode node, edge, and validity maps. Their features are fused into a three-level, width-512 spatial pyramid $F_j^{(\ell)}$. At each level, a temporal transformer encodes consecutive

feature differences $\Delta F_j^{(\ell)} = F_j^{(\ell)} - F_{j-1}^{(\ell)}$, with zero difference at the first step. At selected backbone block b , spatial and temporal residuals jointly update video tokens:

$$\begin{cases} R_{b,j}^s = Z_b^s \left(U_b(F_j^{(\ell_b)}) \right), \\ R_{b,j}^t = Z_b^t \left(U_b(\text{Tr}_g(\Delta F^{(\ell_b)})_j) \right), \\ \tilde{X}_{b,j} = X_{b,j} + R_{b,j}^s + R_{b,j}^t. \end{cases} \quad (4)$$

U_b resizes features to the backbone token grid and Z_b^s, Z_b^t are zero-initialized projections to its hidden dimension. Validity

gates mask unavailable controls, and the reference latent receives no generated-frame residual. In the 28-block backbone, injection occurs before blocks 1, 8, 15, 22, and 28, using pyramid levels 1, 1, 2, 3, and 3, respectively.

b) Global–local kinematic modulation: This path gives projected motion explicit instrument meanings. We derive nine descriptors per instrument from image-plane keypoints: the two-dimensional velocity of P_2 , the angular velocity of the projected shaft direction, projected wrist angle and angular velocity, projected jaw opening angle and angular velocity, and the two-dimensional velocity of the jaw midpoint. Shaft angular velocity measures the change of the P_1P_2 direction in the image plane. Velocities divide coordinate or wrapped angular differences by elapsed time. Each descriptor has a validity indicator; valid values are temporally pooled and standardized using training-set statistics, and invalid entries are set to zero.

The global descriptor group contains translation and shaft-direction change (six values across two instruments); the local group contains wrist and jaw articulation and jaw-midpoint motion (twelve values). Appending validity indicators gives 12- and 24-dimensional inputs, respectively. Separate width-256 temporal encoders process embedded global and local descriptor sequences κ^g and κ^l . Their outputs at latent step j are concatenated as k_j . Let e_σ, m_σ denote noise-level terms and e_j^a, m_j^a the original action terms. Kinematics enters the backbone as

$$\begin{cases} e_j = \text{Norm}(e_\sigma + e_j^a + Z_e(k_j)), \\ m_j = m_\sigma + m_j^a + Z_m(k_j). \end{cases} \quad (5)$$

Here, e_j is the backbone conditioning embedding and m_j is its adaptive layer-normalization modulation input; superscript a denotes the original action pathway. The zero-initialized projections Z_e and Z_m output 2048 and 3×2048 channels. All control temporal encoders use two layers and eight heads. Zero-initialized output projections preserve the pretrained backbone’s function at the start of control training.

D. Teacher Training and Causal Distillation

a) Structurally conditioned teacher: Let z denote clean video latents and let c collect the reference image, actions, and structural conditions. For Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, the interpolated latent is $z_\sigma = (1 - \sigma)z + \sigma\epsilon$. We train the full-sequence teacher with the backbone’s rectified-flow objective:

$$\mathcal{L}_{\text{RF}} = \mathbb{E} \left[w(\sigma) \| v_\theta(z_\sigma, \sigma; c) - (\epsilon - z) \|^2_{\text{mean}} \right]. \quad (6)$$

The notation $\|\cdot\|_{\text{mean}}^2$ denotes mean squared error over tensor elements, and w is the backbone’s noise-level weighting. The reference frame is observed. Predicted conditioning uses a frozen skeleton predictor. The Oracle reference is trained separately with annotated future structure from the same initialization. The injection ablation uses predicted skeletons and retains native action conditioning.

b) Causal student with the same structural interface:

To carry the structural interface into autoregressive video generation, the student retains both injection paths and factorizes the $J = 18$ future video latents as

$$p_\psi(z_{1:J} | z_0, c) = \prod_{j=1}^J p_\psi(z_j | z_0, z_{1:j-1}, c). \quad (7)$$

Here, ψ denotes student parameters and z_0 the reference latent. Cached video attention accesses only preceding generated latents. Structural encoders process the supplied prospective control window c ; the student retrieves features aligned with each latent. Causality here concerns generated video history, with the prospective controls supplied for the prediction window.

We first warm-start the student using teacher ODE trajectories, regressing its clean-latent estimates to the trajectory endpoint on future frames. We then use asymmetric distribution matching, following DMD2 [12] and bidirectional-to-causal distillation [13], with self-generated rollout histories as in Self Forcing [31]. Let y be a student-generated latent sequence. In the distillation parameterization, it is re-noised as $x_\tau = y \cos \tau + \eta \sin \tau$, where η is Gaussian noise. A frozen bidirectional teacher D_T and a trainable bidirectional student-distribution denoiser D_f predict clean latents under the same condition c , yielding $d_T = D_T(x_\tau, \tau; c)$ and $d_f = D_f(x_\tau, \tau; c)$. Define $g = (d_f - d_T)/s$, with $s = \max(\text{mean}|y - d_T|, 10^{-5})$. The generator surrogate is

$$\mathcal{L}_{\text{DMD}} = \mathbb{E} [\|y - \text{sg}(y - g)\|_{\text{mean}}^2]. \quad (8)$$

Here sg stops gradients. The auxiliary denoiser reconstructs detached student samples with $\sin^{-2} \tau$ -weighted squared error. We perform five auxiliary-denoiser updates per generator update and use four denoising steps at inference.

IV. EXPERIMENTS

A. Experimental Setup

We evaluate the action-to-structure-to-video chain on SutureBot and SRTH-Porcine-Cholei, with Cao_cautery_combin for zero-shot video-generator transfer. Zero-shot denotes a video generator evaluated without fine-tuning on Cao; the three-domain predictor benchmark below uses a separate training protocol. SurgCast (Predicted), our main method, uses forecast future skeletons; Oracle supplies ground-truth future skeletons as a reference. The injection ablation also uses predicted skeletons and retains native action conditioning.

SutureBot has 1161/145/146 disjoint training/validation/test episodes and 47,710 complete H73 training windows. Its matched video panel contains 28 eligible validation episodes, each yielding one 72-frame forecast; SRTH teacher comparisons use 95 test episodes and 24 future frames. Both use seed 0, 512×288 resolution, and stride five (approximately 6 Hz for 30 Hz SutureBot recordings).

SutureBot structural teachers share the Cosmos-H DiT initialization and Wan tokenizer, with zero-initialized control

TABLE I

GENERATION QUALITY ON SUTUREBOT AND SRTH-PORCINE-CHOLEI, WITH TEACHERS AND FOUR-STEP STUDENTS GROUPED SEPARATELY.

SURGCAS (PREDICTED) IS THE MAIN METHOD AND USES FORECAST FUTURE SKELETONS; SURGCAS (ORACLE) SUPPLIES GROUND-TRUTH FUTURE SKELETONS AS A REFERENCE CONDITION. THESE LABELS SPECIFY THE SKELETON SOURCE, WHILE THE TEACHER AND STUDENT GROUPS SPECIFY THE VIDEO-GENERATION MODE. THE CAUSAL STUDENT USES EMA-600. GREEN (BOLD) AND YELLOW MARK THE BEST AND SECOND-BEST RESULTS PER GROUP AND DATASET. COSMOS-PREDICT2.5-2B RESULTS ARE FROM [34].

Method	SutureBot							SRTH-Porcine-Cholei						
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	GATC↑	TCD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	GATC↑	TCD↓
<i>Teachers</i>														
Cosmos-Predict2.5-2B [34]	8.1037	0.2249	0.7951	394.0091	6815.0560	0.0036	293.72	10.4087	0.1888	0.7492	458.5920	9802.3112	0.1340	293.7209
Ctrl-World [5]	21.5191	0.7856	0.2321	49.6254	248.3254	0.3656	71.95	19.3131	0.6584	0.3222	88.1025	750.3728	0.3507	70.3725
IRASim [4]	17.9001	0.5556	0.4226	135.8555	1226.6812	0.3551	68.04	16.6724	0.3632	0.4786	218.6194	1564.5107	0.4632	81.2678
Cosmos-H-Surgical-Simulator [6]	22.1164	0.8041	0.1719	31.7472	172.4712	0.6522	47.59	18.6471	0.6387	0.3027	92.6880	698.0070	0.4363	63.5268
SurgCast (Predicted)	24.4687	0.8458	0.1073	27.0716	107.5858	0.7902	40.37	21.3761	0.7530	0.1523	37.0874	228.1920	0.6298	49.7813
SurgCast (Oracle)	25.0237	0.8534	0.1015	26.0625	104.4354	0.8109	38.75	21.4431	0.7546	0.1511	36.7987	222.0160	0.6318	48.1370
<i>Students</i>														
Cosmos-H-Surgical-Simulator (Original Student)	22.1901	0.7908	0.1943	59.0538	212.5411	0.5151	60.36	17.5261	0.5932	0.4224	92.3297	700.6836	0.3236	82.0362
SurgCast (Causal Student)	23.2639	0.8114	0.1504	50.1808	191.5575	0.6278	47.50	20.2579	0.7134	0.1753	49.7365	315.0477	0.5411	53.7209

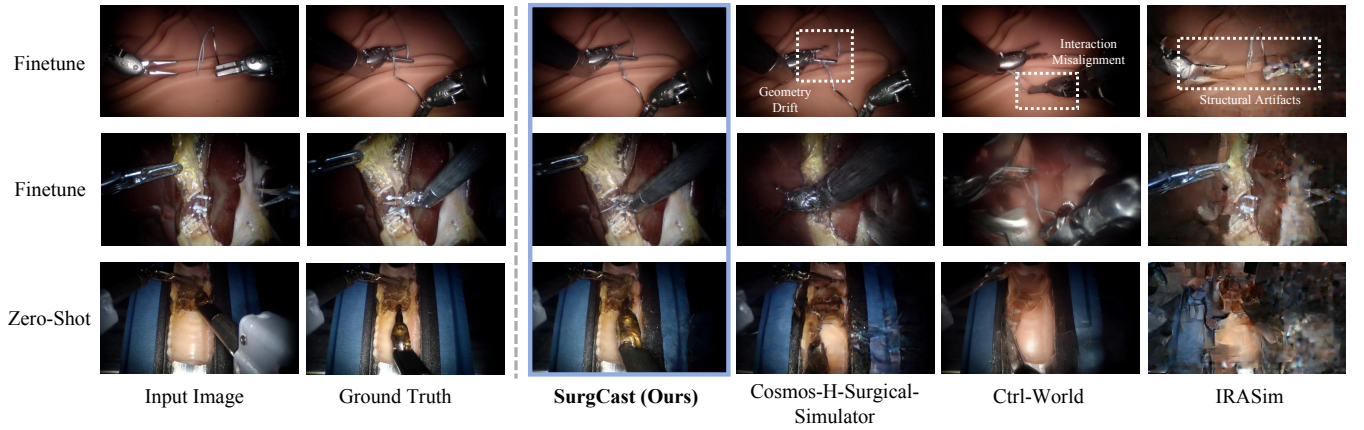


Fig. 3. Teacher outputs on SutureBot (top), SRTH-Porcine-Cholei (middle), and Cao.cautery_combin with zero-shot video-generator transfer (bottom). The figure label 'SurgCast (Ours)' denotes SurgCast (Predicted), conditioned on predicted future skeletons. White dashed annotations identify geometry drift, interaction misalignment, and structural artifacts in baseline outputs; compare the annotated instrument regions with SurgCast under the corresponding scene and action sequence.

outputs. Training uses four H100 GPUs, effective batch 64, and up to 1600 AdamW updates (learning rate 4×10^{-5} , weight decay 0.1). Teachers use 35 sampling steps with guidance 0. Ctrl-World and IRASim use 50 steps with guidance 1 and native action/temporal interfaces, adapted on the same SutureBot training partition. SutureBot current-frame keypoints, visibility, confidence, and boxes come from offline framewise SurgPose detections; robot state comes from the current observation and tool identities from episode metadata. A frozen predictor forecasts 72 steps jointly from these current inputs and future actions.

PSNR, SSIM, and LPIPS assess frame fidelity; FID and FVD assess image- and video-level quality. FVD uses the first 16 future frames in the SutureBot and SRTH comparisons. GT-Anchored Tool Consistency (GATC) compares grayscale appearance in ground-truth tool regions using zero-mean normalized cross-correlation, a gradient-presence penalty, and local alignment within ± 3 pixels and ± 1 frame. Tool Centroid Distance (TCD) matches Medical-SAM3 tool-mask centroids by Hungarian assignment; a missing tool incurs half the image diagonal. Both use medians over valid future frames and exclude the observed first frame. For

rollout evaluation, C1–C6 partition the 72 future frames into six consecutive 12-frame segments. FDS-L1 and FDS-SSIM report segment-averaged framewise L1 error and SSIM against the ground-truth video, with RGB values normalized to $[-1, 1]$; GATC and TCD are also reported by segment.

Table III macro-averages H12 prediction over SutureBot, SRTH-Porcine-Cholei, and Cao.cautery_combin, each with 1161/145/146 training/validation/test windows from disjoint episodes. It evaluates a jointly trained model without adaptation and separate domain-adapted models over 12 future steps at stride five. ADE and FDE measure mean and final-step error in 960×540 coordinates; PCK@20px is the fraction within 20 pixels. The 72-step video controls use a separately trained predictor. Throughput and peak GPU memory are measured over 10 four-step, 73-frame rollouts.

B. Generation Quality and Instrument Consistency

We first ask whether the predicted structural interface improves action-conditioned video generation. Table I compares SurgCast (Predicted), our main method, with the direct-action teacher. It improves all seven reported metrics on both datasets. FVD decreases by 37.6% on SutureBot and

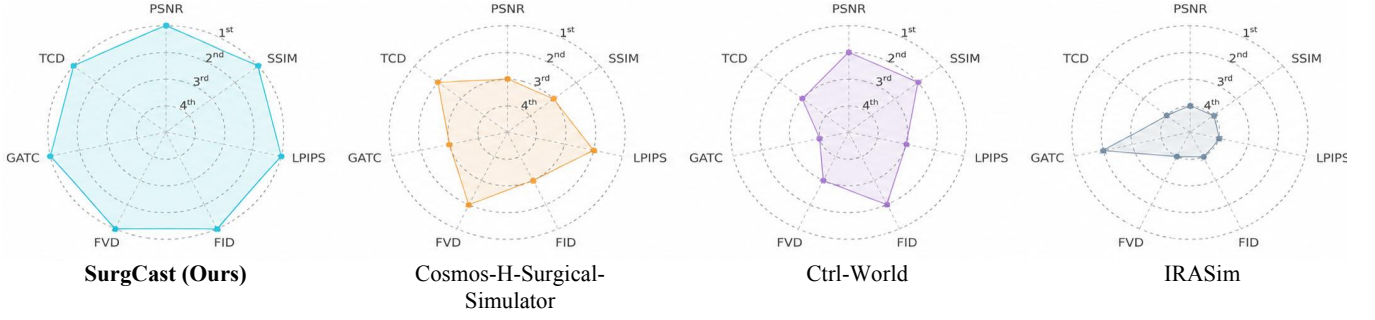


Fig. 4. Video-generator transfer to Cao.cautery.combin without video-model fine-tuning: ranks across seven quality and instrument-consistency metrics. ‘SurgCast (Ours)’ labels the SurgCast teacher; the outer ring denotes rank 1. The radial positions encode ranking within each metric, not metric magnitudes.

67.3% on SRT-H-Porcine-Cholei, alongside improved GATC and TCD. The paired results show improved video quality together with closer instrument appearance and position under the supplied actions.

The separately trained Oracle reference uses ground-truth future structure. Predicted and Oracle yield FVD values of 107.59 and 104.44 on SutureBot, and 228.19 and 222.02 on SRT-H-Porcine-Cholei. Relative to the gains over direct-action conditioning, the modest differences show that predicted structure retains much of the generation benefit observed with annotated structure.

The four-step causal student tests whether the interface carries through distillation. It improves all seven metrics over the original student, with FVD decreasing from 212.54 to 191.56 on SutureBot and from 700.68 to 315.05 on SRT-H-Porcine-Cholei. This comparison extends the full-sequence result to causal generation under the same four-step student sampling budget.

Figure 3 illustrates how the predicted-skeleton teacher preserves instrument structure against baseline geometry drift, interaction misalignment, and structural artifacts. These examples complement the Predicted results in Table I.

C. Ablation Studies

The ablations examine both links in the structural interface: how geometric and kinematic injection use the forecast, and what prospective commands contribute to skeleton prediction.

a) Dual-path semantic-kinematic injection: Table II tests the two injection paths under predicted skeletons in the SutureBot teacher. The action-only control uses the same training recipe; all variants retain native action conditioning. Either structural path improves all six metrics over action-only. Geometry yields lower LPIPS and FID, while kinematics yields higher PSNR, SSIM, and GATC and lower TCD. Combining both paths performs best on every metric, reaching PSNR of 24.4687, LPIPS of 0.1073, and TCD of 40.37. The combined row is SurgCast (Predicted) in Table I. These results support complementary use of geometric residuals and derived motion semantics in the main predicted-skeleton generation pathway.

b) Action-to-future-skeleton prediction: Table III tests what future actions add to current structure and robot state. For the joint three-domain predictor, actions reduce ADE

TABLE II
PREDICTED-SKELETON ABLATION IN THE SUTUREBOT TEACHER.
ENABLED PATHS USE FORECAST SKELETONS; ALL VARIANTS, INCLUDING ACTION-ONLY, USE THE SAME TRAINING RECIPE AND RETAIN NATIVE ACTION CONDITIONING. YELLOW MARKS THE BEST SINGLE-PATH RESULT; GREEN MARKS THE COMBINED MODEL.

Method	Geo.	Kin.	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	GATC $\uparrow$	TCD $\downarrow$
Action only	—	—	22.1164	0.8041	0.1719	31.7472	0.6522	47.59
Geometry only	✓	—	23.0351	0.8337	0.1212	27.8214	0.7815	41.7564
Kinematics only	—	✓	23.5359	0.8368	0.1370	28.8493	0.7869	41.3922
Geometry + Kinematics	✓	✓	24.4687	0.8458	0.1073	27.0716	0.7902	40.37

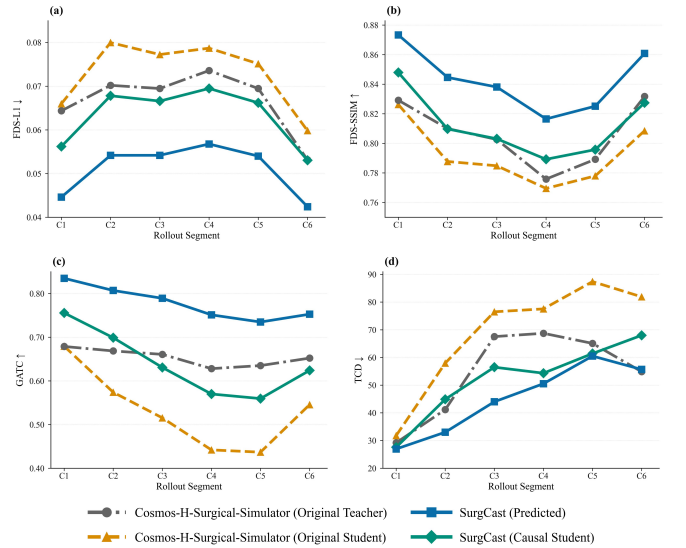


Fig. 5. Segment-wise rollout on 28 SutureBot validation episodes. C1–C6 cover successive 12-frame portions of the 72-frame prediction. Blue squares denote SurgCast (Predicted), conditioned on predicted future skeletons; student curves compare four-step causal generators. Panels (a,b) report frame fidelity; (c,d) report tool-region consistency and centroid distance.

from 29.30 to 15.94 and FDE@H12 from 41.33 to 20.03, while PCK@20px rises from 57.65% to 75.08%. All four metrics also improve over persistence.

Domain-adapted H12 predictors improve ADE and PCK in every interval. At steps 9–12, actions reduce ADE from 38.82 to 19.07 and increase PCK from 48.06% to 68.45% over prediction without future actions.

The 72-step control predictor is separately evaluated on

TABLE III

H12 SKELETON PREDICTION: THREE-DOMAIN MACRO-AVERAGE IN 960×540 COORDINATES. TOP: JOINTLY TRAINED PREDICTOR WITHOUT FURTHER ADAPTATION. BOTTOM: DOMAIN-ADAPTED PREDICTORS OVER STEPS 1–4, 5–8, AND 9–12 (STRIDE 5). INPUTS CUMULATIVELY ADD r_0 AND $a_{1:H}$ TO S_0 . YELLOW MARKS THE BEST BASELINE; GREEN MARKS THE ACTION-CONDITIONED PREDICTOR.

Overall Performance						
Method	Input	ADE↓	FDE@H12↓	PCK@20px↑	Vis. Acc.↑	
Persistence	S_0	55.64	80.83	49.04%	92.96%	
Predictor w/o future action	$+r_0$	29.30	41.33	57.65%	97.33%	
Action-to-Skeleton Predictor	$+a_{1:H}$	15.94	20.03	75.08%	97.89%	
Performance by Horizon						
Method	ADE↓			PCK@20px↑		
	Early	Mid	Late	Early	Mid	Late
Persistence	30.43	59.95	76.71	62.98%	44.66%	39.36%
Predictor w/o future action	18.34	30.87	38.82	71.18%	53.58%	48.06%
Action-to-Skeleton Predictor	12.42	16.38	19.07	82.77%	73.91%	68.45%

732 windows from 28 SutureBot validation episodes using 960×540 pose pseudo-labels. Its valid-point ADE is 30.76 pixels, versus 96.07 for a separately trained no-action model; over steps 37–72, the errors are 32.38 and 119.59 pixels. This supports action-conditioned forecasting at the full video horizon.

D. Temporal Consistency and Video-Generator Transfer

a) Temporal rollout: Figure 5 examines structural conditioning across six SutureBot rollout segments. SurgCast (Predicted) has lower FDS-L1 and higher FDS-SSIM and GATC than the direct-action teacher in C1–C6; its TCD is lower in the first five segments and slightly higher at C6.

The causal student improves all four reported metrics over the original student in every segment. These within-segment gains show that the structural interface remains beneficial across the evaluated rollout, while the nonmonotonic curves reflect variation between segments.

b) Zero-shot video-generator transfer: On Cao_cautery_combin, without video-generator fine-tuning, Fig. 4 ranks the SurgCast teacher against Cosmos-H-Surgical-Simulator, Ctrl-World, and IRASim. SurgCast ranks first on PSNR, SSIM, LPIPS, FID, FVD, GATC, and TCD. The bottom row of Fig. 3 shows the corresponding visual comparison. Together, the ranks and example extend the evaluated generation benefit to this video-generator transfer setting.

E. Operator-Driven Interaction

Figure 7 qualitatively tests action responsiveness from a fixed initial image and instrument skeleton. Distinct 6-DoF command trajectories produce different predicted skeletons



Fig. 6. Operator-driven SurgCast with a physical 6-DoF surgical master device (left) and VR controllers (right) as sources of prospective action commands. Displays show the generated surgical view; action-dependent outputs are compared in Fig. 7.

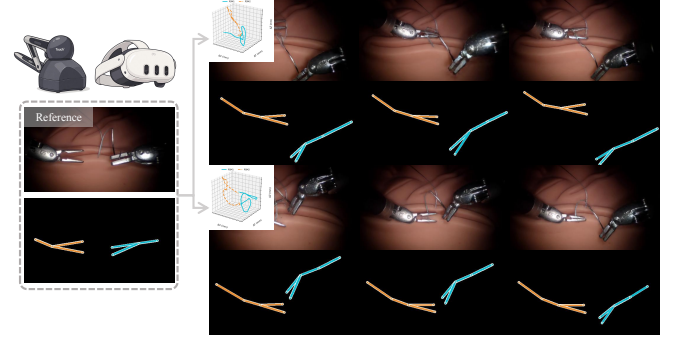


Fig. 7. Action branching from the same initial image and skeleton (left). Distinct 6-DoF commands (insets) produce different predicted skeletons and video rollouts (upper and lower rows).

and generated instrument motions. The commands can be supplied through a physical surgical master device or a companion VR interface (Fig. 6).

Across 10 four-step, 73-frame rollouts, the causal student achieves 10.36 FPS in rollout throughput with 8.703 GB peak GPU memory, compared with 10.57 FPS and 8.481 GB for the original student.

V. CONCLUSION

SurgCast establishes predicted instrument structure as an explicit interface between prospective robot actions and video generation. Its two injection paths expose the same forecast as image-aligned geometry and instrument-meaningful motion descriptors. Across the evaluated datasets, this interface improves video fidelity and instrument consistency over direct-action baselines. Ablations show that future commands improve structural forecasting and that both injection paths contribute under predicted skeletons; their combination performs best across the six SutureBot ablation metrics. The gains also persist in four-step causal generation. Together, these findings support forecasting articulated structure and explicitly using its geometry and motion semantics for surgical video control.

The evaluation uses supplied action windows; sustained closed-loop robot control remains to be validated. Testing the interface under continuous feedback is the next step toward robotic decision-making use.

REFERENCES

- [1] C. Finn and S. Levine, “Deep visual foresight for planning robot motion,” in *IEEE International Conference on Robotics and Automation*, 2017.
- [2] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. B. Tenenbaum, D. Schuurmans, and P. Abbeel, “Learning universal policies via text-guided video generation,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [3] S. Yang, Y. Du, S. Ghasemipour, J. Tompson, L. P. Kaelbling, D. Schuurmans, and P. Abbeel, “Learning interactive real-world simulators,” in *International Conference on Learning Representations*, 2024.
- [4] F. Zhu, H. Wu, S. Guo, Y. Liu, C. Cheang, and T. Kong, “IRASim: A fine-grained world model for robot manipulation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 9834–9844.
- [5] Y. Guo, L. X. Shi, J. Chen, and C. Finn, “Ctrl-World: A controllable generative world model for robot manipulation,” in *International Conference on Learning Representations*, 2026.
- [6] J. G. Tejero, L. Zbinden, K. Sheth, R. K. M. N. Daher, D. G. M. na, F. Binkiewicz, P. Thornycroft, M. Azizian, and S. D. Huver, “NVIDIA Cosmos-H-Dreams: Real-time generative physics simulation for surgical robotics,” *arXiv preprint arXiv:2608.24199*, 2026.
- [7] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, “VideoComposer: Compositional video synthesis with motion controllability,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [8] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, “DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory,” *arXiv preprint arXiv:2308.08089*, 2023.
- [9] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, “MotionCtrl: A unified and flexible motion controller for video generation,” in *ACM SIGGRAPH Conference Papers*, 2024, pp. 1–11.
- [10] S. Rapuri, L. Seenivasan, D. Schneider, R. Soberanis-Mukul, Y. He, H. Ding, J. Xu, C. Yu, C. Jing, P. Guo, D. Xu, and M. Unberath, “SAW: Toward a surgical action world model via controllable and scalable video generation,” *arXiv preprint arXiv:2603.13024*, 2026.
- [11] B. Li, S. Yang, B. Peng, X. Guo, E. Zhang, Y. Tao, J. Duan, D. Xu, Q. Dou, X. Jin, W. Zeng, H. Zhao, and Y. Jin, “From articulated kinematics to routed visual control for action-conditioned surgical video generation,” *arXiv preprint arXiv:2605.08712*, 2026.
- [12] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and W. T. Freeman, “Improved distribution matching distillation for fast image synthesis,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [13] T. Yin, Q. Zhang, R. Zhang, W. T. Freeman, F. Durand, E. Shechtman, and X. Huang, “From slow bidirectional to fast autoregressive video diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [14] J. Wu, S. Yin, N. Feng, X. He, D. Li, J. Hao, and M. Long, “iVideoGPT: Interactive VideoGPTs are scalable world models,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [15] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, “DINO-WM: World models on pre-trained visual features enable zero-shot planning,” in *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267, 2025, pp. 79 115–79 135.
- [16] L. Wang, K. Zhao, C. Liu, and X. Chen, “Learning real-world action-video dynamics with heterogeneous masked autoregression,” *arXiv preprint arXiv:2502.04296*, 2025.
- [17] Open-H-Embodiment Consortium, “Open-h-embodiment: A large-scale dataset for enabling foundation models in medical robotics,” *arXiv preprint arXiv:2604.21017*, 2026.
- [18] M. Yoshimura, M. M. Marinho, K. Harada, and M. Mitsuishi, “MBA-Pose: Mask and bounding-box aware pose estimation of surgical instruments with photorealistic domain randomization,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 9445–9452.
- [19] Z. Wu, A. Schmidt, R. Moore, H. Zhou, A. Banks, P. Kazanzides, and S. E. Salcudean, “SurgPose: A dataset for articulated robotic surgical tool pose estimation and tracking,” in *IEEE International Conference on Robotics and Automation*, 2025.
- [20] B. Ghanekar, L. R. Johnson, J. L. Laughlin, M. K. O’Malley, and A. Veeraraghavan, “Video-based surgical tool-tip and keypoint tracking using multi-frame context-driven deep learning models,” in *IEEE International Symposium on Biomedical Imaging*, 2025.
- [21] Y. Qin, S. Feyzabadi, M. Allan, J. W. Burdick, and M. Azizian, “daVinciNet: Joint prediction of motion and surgical state in robot-assisted surgery,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.
- [22] C. Shi, Y. Zheng, and A. M. Fey, “Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022.
- [23] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [24] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, “T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4296–4304.
- [25] Y. Yeganeh, R. Lazuardi, A. Shamseddin, E. Dari, Y. Thirani, N. Navab, and A. Farshad, “VISAGE: Video synthesis using action graphs for surgery,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*, 2025, pp. 146–156.
- [26] S. Koju, S. Bastola, P. Shrestha, S. Amgain, Y. R. Shrestha, R. P. K. Poudel, and B. Bhattarai, “Surgical vision world model,” in *Data Engineering in Medical Imaging*, 2026, pp. 1–10.
- [27] T. Chen, S. Yang, J. Wang, L. Bai, H. Ren, and L. Zhou, “SurgSora: Object-aware diffusion model for controllable surgical video generation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15969, 2026, pp. 521–531.
- [28] T. Salimans and J. Ho, “Progressive distillation for fast sampling of diffusion models,” in *International Conference on Learning Representations*, 2022.
- [29] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, “One-step diffusion with distribution matching distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6613–6623.
- [30] B. Chen, D. M. Monso, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann, “Diffusion forcing: Next-token prediction meets full-sequence diffusion,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [31] X. Huang, Z. Li, G. He, M. Zhou, and E. Shechtman, “Self forcing: Bridging the train-test gap in autoregressive video diffusion,” in *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [32] K. Liu, W. Hu, J. Xu, Y. Shan, and S. Lu, “Rolling forcing: Autoregressive long video diffusion in real time,” in *International Conference on Learning Representations*, 2026, pp. 91 177–91 196.
- [33] H. Zhu, M. Zhao, G. He, H. Su, C. Li, and J. Zhu, “Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation,” in *International Conference on Machine Learning*, 2026.
- [34] NVIDIA, A. Ali, J. Bai, M. Bala, Y. Balaji, A. Blakeman, T. Cai, J. Cao, T. Cao, E. Cha, Y.-W. Chao, P. Chattopadhyay, M. Chen, Y. Chen, Y. Chen, S. Cheng, Y. Cui, J. Diamond, Y. Ding, J. Fan, L. Fan, L. Feng, F. Ferroni, S. Fidler, X. Fu, R. Gao, Y. Ge, J. Gu, A. Gupta, S. Gururani, I. El Hanafi, A. Hassani, Z. Hao, J. Huffman, J. Jang, P. Jannaty, J. Kautz, G. Lam, X. Li, Z. Li, M. Liao, C.-H. Lin, T.-Y. Lin, Y.-C. Lin, H. Ling, M.-Y. Liu, X. Liu, Y. Lu, A. Luo, Q. Ma, H. Mao, K. Mo, S. Nah, Y. Narang, A. Panaskar, L. Pavao, T. Pham, M. Ramezani, F. Reda, S. Reed, X. Ren, H. Shao, Y. Shen, S. Shi, S. Song, B. Stefaniak, S. Sun, S. Tang, S. Tasmeen, L. Tchapti, W.-C. Tseng, J. Varghese, A. Z. Wang, H. Wang, H. Wang, H. Wang, T.-C. Wang, F. Wei, J. Xu, D. Yang, X. Yang, H. Ye, S. Ye, X. Zeng, J. Zhang, Q. Zhang, K. Zheng, A. Zhu, and Y. Zhu, “World simulation with video foundation models for physical ai,” *arXiv preprint arXiv:2511.00062*, 2025.