
Physics-Conditioned Video Diffusion with Kinematic Priors for Fusion Capsule Polishing

Shashank Galla¹, Abhishek Hanchate¹, Monika Biener²,
Suhas Bhandarkar², Satish Bukkapatnam¹

¹Texas A&M University, College Station, TX, USA

²Lawrence Livermore National Laboratory, Livermore, CA, USA

Abstract

Video diffusion models can generate realistic process videos, but using them as controllable surrogates for industrial systems remains difficult when key control variables are only partially observable and no simulator or differentiable physics model is available. We study this regime in inertial confinement fusion (ICF) target capsule polishing, where surface defects can degrade fusion yield. The dynamics are governed by the polishing speed ω_p and capsule-pad slip S_C . While capsule motion can be monitored from video, S_C is only indirectly observable. Using only an analytical kinematic model, we adapt a frozen Stable Video Diffusion (SVD-xt) backbone via LoRA and condition the generation on Fourier-encoded physics parameters (ω_p, S_C) . Two training objectives operate on latent frame-difference dynamics: (i) a batch level correlation loss aligning them with the analytical sliding speed, and (ii) a stratified ratio-matching loss that, conditioned on ω_p , isolates slip structure from the dominant ω_p -driven variance. We evaluate **kinematic** consistency by tracking capsule motion in generated videos, inverting the kinematic model, and measuring rank agreement under a physical validity gate. For speed, rank accuracy exceeds 0.97 on real data; for slip, it reaches 0.735 with Spearman ≥ 0.50 on synthetic data. Overall, the results suggest a practical recipe for controllable industrial video surrogates when only low-dimensional analytical kinematics are available.

1 Introduction

Inertial confinement fusion (ICF) target capsules must be polished to nanometer tolerances because defects of only tens of nanometers can seed hydrodynamic instabilities and degrade ignition symmetry [23, 16, 1]. Polishing is therefore a costly throughput bottleneck, and improving it requires understanding how operating speed ω_p and capsule-pad slip S_C drive material removal. These dynamics are only partly observable: the capsule moves stochastically inside an opaque fixture, and routine instrumentation consists of a top-down video stream.

We seek *controllable video surrogates*: given a real reference frame and prescribed (ω_p, S_C) , the model samples plausible capsule-motion clips. Sampling supports counterfactual sweeps, uncertainty estimates for derived kinematics, and anomaly detection from the distribution of plausible outcomes. Analytical polishing kinematics [19, 20] also allows generated clips to be queried in physical coordinates: circulation, spin, and sliding speed. Most physics-guided video generation assumes PDE residuals, calibrated simulators, trajectory priors, dense motion supervision, or post-training guidance [3, 15, 39, 21, 43, 22, 34, 2, 37, 40]. Capsule polishing has none of these: the state is partly hidden, no calibrated video simulator exists for real footage, and the available physics is only a low-dimensional analytical map from (ω_p, S_C) to kinematic scalars. We therefore combine a stochastic video prior trained on real footage with scalar kinematic constraints.

Our approach adapts a frozen Stable Video Diffusion (SVD-xt) [4] image-to-video backbone with LoRA. Fourier tokens encode (ω_p, S_C) and enter cross-attention. The kinematics define two weak supervision signals without absolute labels: a batch correlation loss $\mathcal{L}_{\text{dyn}}$ aligning latent frame differences with analytical sliding speed, and a stratified ratio loss $\mathcal{L}_{\text{ratio}}$ matching within- ω_p latent dynamics ratios to the analytical slip ratio. Because real S_C is unobserved, we treat it as a prior-regularized latent conditioning variable and measure controllability by tracking generated clips and inverting them back to kinematics.

We combine these standard components to study controllable video generation under the sensing constraints of ICF capsule polishing. Our contributions are:

(i) Video generation with analytical kinematic supervision. We formulate a video surrogate that combines scalar conditioning with two latent objectives for analytical speed and slip relations. The formulation uses process videos and analytical kinematics without requiring a calibrated simulator of the real process or dense motion labels.

(ii) Kinematic evaluation of generated motion. We recover capsule circulation from generated pixels and use kinematic inversion to assess speed and slip ordering. We report rank accuracy, confidence intervals, and the fraction of clips passing validity checks alongside visual metrics. These scores measure consistency with the analytical model shared by training and evaluation.

(iii) Speed control and the limits of slip recovery. We distinguish observable speed control in real footage from slip ordering in synthetic clips with known labels. Across five evaluation seeds, the full model reaches real speed rank 0.985 and synthetic slip rank 0.735 (Tables 1 and 2). Real slip ordering remains at chance. Ablations show that scalar conditioning supplies most of the ordering, while the combined losses reduce synthetic FID-I3D and FVD without improving slip rank over conditioning alone.

2 Related Work

Video diffusion and scalar conditioning. Latent and video diffusion models provide strong short-horizon visual priors [26, 13, 4, 24], with LoRA enabling adaptation to small industrial datasets [14]. Continuous process controls require mechanisms that distinguish nearby physical values; common choices include FiLM, cross attention, ControlNet or IP-Adapter branches, and Fourier features [25, 41, 38, 29].

Physics guided video generation. Recent methods inject physics as field residuals or score guidance for known PDEs [3, 15], simulators for rigid bodies, materials, or interactions [39, 21, 5, 43, 34, 2], or dense optical-flow/trajectory signals [22, 35, 45]. Object-centric methods similarly assume spatial descriptors such as masks, contact fields, or skeleton tracks. Capsule polishing lacks these signals: the state is hidden, dense correspondences are unreliable on a featureless specular shell, and the only physics is the scalar pair (ω_p, S_C) . We compare against flow and analytical warp baselines in physical coordinates (Sec. 6.5); dense-flow or trajectory-token baselines require supervision that is unavailable here.

LoRA prediction and hypernetwork control. Recent controllers can alter model weights or use scaling tokens: TC-LoRA predicts time-varying LoRA weights, Video2LoRA predicts LoRA modules from a reference video [6, 36], and LiON-LoRA-style fusion designs use tokens to modulate motion strength. Our controls are physical scalars supervised by kinematic ordering, so we keep one LoRA adapter set and inject Fourier operating tokens at every cross-attention layer. This gives a simple conditioning-only baseline analogous to an inference-time scaling-token controller: the “**conditioning only**” rows keep the tokens but remove $\mathcal{L}_{\text{dyn}}, \mathcal{L}_{\text{ratio}}$, isolating the marginal effect of the losses.

Kinematic-consistency evaluation. General video benchmarks for physical plausibility and scientific diffusion emulators emphasize that perceptual realism is not enough for physical use. Our metric follows that principle at the domain scale: instead of reporting only FID/FVD, we recover $\hat{\omega}_c$ and $\hat{R}_T$ from generated pixels and score monotonic agreement with operating variables. This does not make the videos mechanically exact, but it prevents appearance metrics from being mistaken for process controllability.

3 Preliminaries

Problem. Given a dataset $\mathcal{D} = \{(\mathbf{v}_i, \omega_p^{(i)}, S_C^{(i)}, \mathbf{x}_{\text{cond}}^{(i)})\}_{i=1}^N$ of clips $\mathbf{v}_i \in \mathbb{R}^{T \times 3 \times H \times W}$, baseplate speed $\omega_p^{(i)}$, head-contact slip $S_C^{(i)}$, and first frame $\mathbf{x}_{\text{cond}}^{(i)}$, we learn $p_\theta(\mathbf{v} \mid \omega_p, S_C, \mathbf{x}_{\text{cond}})$ which clips are visually realistic and whose dynamics scale with $V_{\text{eff}}(\omega_p, S_C)$ (speed) and $R_T(S_C)$ (slip).

Analytical kinematics.

Following Lee’s kinematic model [19, 20], a rigid ball of radius R_b rests in a V-groove of half-angle θ and radius R_g on a baseplate rotating at ω_p . No slip at the two groove contacts and head-contact slip $S_C \in [0, 1]$ give the scalar map below (derivation in Appendix A1); Fig. 1 shows the contacts.

Two facts drive the losses: for fixed slip, V_{eff} is linear in ω_p with zero intercept; for $\theta \in (0, \pi/2)$, R_T is strictly increasing in S_C and provides an ordinal slip target.

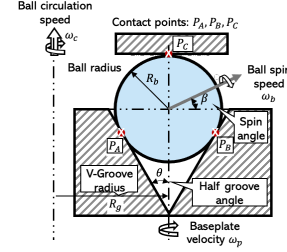


Figure 1: Polishing setup.

$$\omega_c = \omega_p \frac{1 + S_C \sin \theta}{1 + \sin \theta}, \quad \omega_b = \frac{R_g \omega_p (1 - S_C)}{R_b \cos \beta (1 + \sin \theta)}, \quad V_C = S_C R_g \omega_p, \quad (1)$$

where ω_c is shell circulation, ω_b is self-spin, V_C is head-contact sliding speed, and $\cos \beta = R_g / \sqrt{R_g^2 + R_b^2 \sin^2 \theta}$ it is set by geometry. We define

$$V_{\text{eff}}(\omega_p, S_C) = \sqrt{(R_g \omega_c)^2 + (R_b \omega_b)^2 + V_C^2}, \quad R_T(S_C) \triangleq \frac{\omega_c}{\omega_p} = \frac{1 + S_C \sin \theta}{1 + \sin \theta}. \quad (2)$$

Our apparatus has $\theta=45^\circ$ and speeds $\{21, 48, 78, 108\}$ RPM. Real clips use shells bonded to the baseplate, so the head contact is nominally pure sliding ($S_C=1$) [19]; compliant-pad deformation, pressure gradients, and fluid effects can lower the effective slip [17], but per-clip S_C is unobservable from overhead video. We therefore use $S_C \sim \text{Beta}(199.5, 10.5)$ (mean 0.95), rescaled to $[0.92, 0.98]$, as an uncertainty prior and regularizer, not a measured label [9]. To test slip when labels exist, we also build synthetic clips with $S_C \in \{0.20, 0.40, 0.60, 0.80\}$ (Sec. 6).

4 Method

The framework has three parts: a frozen video backbone with LoRA adapters, Fourier physics tokens injected by cross attention, and latent losses computed from differentiable kinematics.

Backbone. We adopt SVD-xt [4], which operates on VAE latents $\mathbf{z} \in \mathbb{R}^{T \times C \times h \times w}$ with v-prediction $\mathbf{v}_\theta = \sqrt{\bar{\alpha}_t} \epsilon - \sqrt{1 - \bar{\alpha}_t} \mathbf{x}_0$ and clean-latent recovery $\hat{\mathbf{x}}_0 = \sqrt{\bar{\alpha}_t} \mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \mathbf{v}_\theta$. We freeze the backbone, VAE, and CLIP encoder, and train rank-32 LoRA adapters [14] on attention projections with Min-SNR- γ weighting ($\gamma=5$) [10].

Fourier physics tokens. RPM is normalized to $\hat{\omega}_p = (\omega_p - \omega_{p,\min}) / (\omega_{p,\max} - \omega_{p,\min}) \in [0, 1]$ over $\{21, 48, 78, 108\}$ RPM; S_C is already fractional. Each scalar s is lifted to $\gamma(s) = [\sin(2^k \pi s), \cos(2^k \pi s)]_{k=0}^{K-1}$ and encoded by an MLP into $\mathbb{R}^d$ ($d=1024$). The continuity of γ supports interpolation to unseen speeds, tested by leave-one-speed-out evaluation (Sec. 6); it also permits off-grid queries such as 55 RPM, though we only validate held-out grid speeds here. Tokens $\mathbf{T}_{\text{rpm}}$, $\mathbf{T}_{S_C}$ are combined by multi-head attention with the mean-token query, and the conditioning sequence is

$$\mathbf{T}_{\text{int}} = \text{LN}(\bar{\mathbf{T}} + \text{MHA}(\bar{\mathbf{T}}, \mathbf{T}_{12}, \mathbf{T}_{12})), \quad \mathbf{E} = [\mathbf{e}_{\text{img}}; \mathbf{T}_{\text{rpm}}; \mathbf{T}_{S_C}; \mathbf{T}_{\text{int}}] \in \mathbb{R}^{4 \times d}, \quad (3)$$

with $\mathbf{T}_{12} = [\mathbf{T}_{\text{rpm}}; \mathbf{T}_{S_C}]$ and $\bar{\mathbf{T}} = \frac{1}{2}(\mathbf{T}_{\text{rpm}} + \mathbf{T}_{S_C})$. $\mathbf{E}$ enters every UNet cross-attention layer through LoRA-adapted K, V projections; Q stays frozen to reduce query drift. Self-attention Q, K, V and output projections also receive LoRA adapters ($r=\alpha=32$) across temporal and spatial attention. Image and physics tokens are jointly dropped with $p_{\text{cfg}}=0.1$ for classifier-free guidance [12]. Encoding and mechanism ablations are in Sec. 6: **FiLM replaces cross attention with a feature-wise affine modulation of the UNet activations, and raw S_C keeps cross attention but removes the Fourier lift γ for S_C , feeding the scalar directly.**

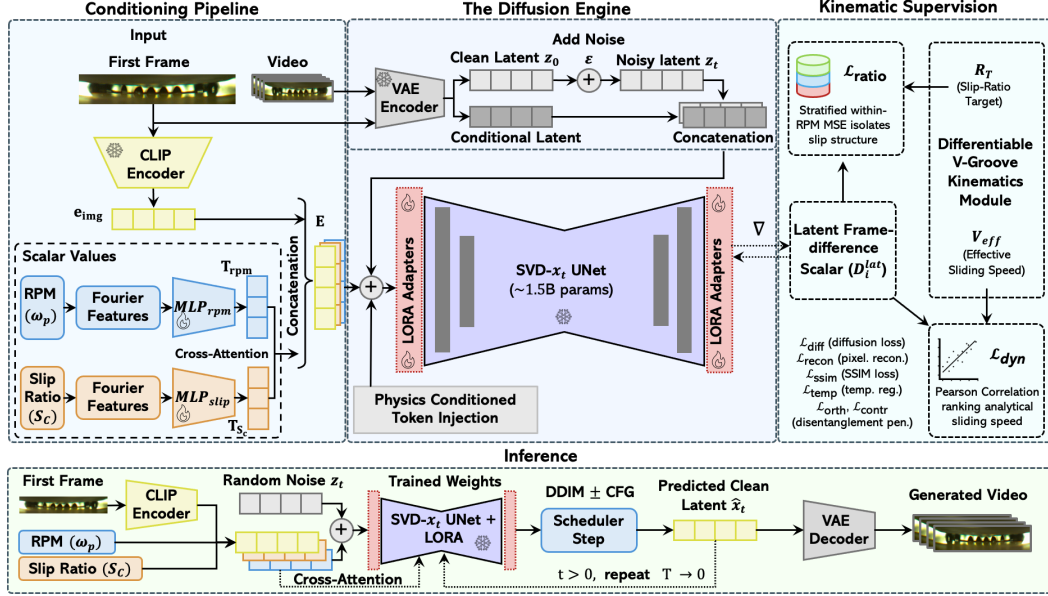


Figure 2: Method overview. SVD-xt is frozen (gray); only LoRA adapters and the Fourier physics encoder (blue) are trained. The operating scalars (ω_p, S_C) become cross-attention tokens and kinematic targets V_{eff}, R_T for latent losses $\mathcal{L}_{\text{dyn}}, \mathcal{L}_{\text{ratio}}$ on $\hat{\mathbf{x}}_0$. **At inference (bottom row) there is no target clip: starting from random noise, the same frozen encoders and trained adapters are reused to alternate the UNet with the sampling scheduler, and the frozen VAE decodes only the final latent to pixels.**

Differentiable kinematics module. Equations (1)–(2) map physical (ω_p, S_C) to $V_{\text{eff}}, \omega_c, \omega_b, R_T$. These are scalar targets only: gradients enter LoRA through the latent frame-difference scalar D^{lat} computed on $\hat{\mathbf{x}}_0$, not through the kinematic constants. Kinematics use physical RPM; only tokens use normalized RPM.

Dynamics consistency loss $\mathcal{L}_{\text{dyn}}$. From $\hat{\mathbf{x}}_0$ we form a per-clip latent frame-difference scalar

$$D_i^{\text{lat}} = \frac{1}{(T-1)Chw} \sum_{t=1}^{T-1} \|\hat{\mathbf{x}}_{0,i}^{(t+1)} - \hat{\mathbf{x}}_{0,i}^{(t)}\|_1, \quad \mathcal{L}_{\text{dyn}} = -\text{corr}(V_{\text{eff}}^{(1:B)}, D^{\text{lat}(1:B)}). \quad (4)$$

Unlike PDE-residual losses [3] that constrain each sample, $\mathcal{L}_{\text{dyn}}$ enforces a *ranking* across samples consistent with the analytical sliding speed.

Stratified velocity-ratio loss $\mathcal{L}_{\text{ratio}}$. Naively replacing V_{eff} with $R_T(S_C)$ in Eq. (4) fails because between-RPM dynamics dominate, while the within-RPM variance of $R_T(S_C)$ under the real Beta prior is only $\text{Var}(R_T | r) \approx 3.85 \times 10^{-5}$. A pooled correlation, therefore, sees mostly RPM. so we stratify by RPM. This also explains why the loss is a weak signal at $B \approx 8-16$: it regularizes ordinal structure rather than providing dense per-sample supervision. For each stratum $\mathcal{B}_r$ in the batch, define mean-normalized signals $\tilde{D}^{(i)} = D_i^{\text{lat}} / \text{sg}[\overline{D^{\text{lat}}_{\mathcal{B}_r}}]$, $\tilde{R}_T^{(i)} = R_T(S_C^{(i)}) / \overline{R_T}_{\mathcal{B}_r}$, where $\overline{D^{\text{lat}}}_{\mathcal{B}_r} = \frac{1}{|\mathcal{B}_r|} \sum_{j \in \mathcal{B}_r} D_j^{\text{lat}}$ and analogously for $\overline{R_T}_{\mathcal{B}_r}$. The stop-gradient $\text{sg}[\cdot]$ ($\cdot \text{.detach}()$) on the stratum-mean denominator prevents the loss from collapsing the predicted ratios to a degenerate scale. We use

$$\mathcal{L}_{\text{ratio}} = \frac{1}{|\mathcal{R}_B|} \sum_{r \in \mathcal{R}_B} \frac{1}{|\mathcal{B}_r|} \sum_{i \in \mathcal{B}_r} (\tilde{D}^{(i)} - \tilde{R}_T^{(i)})^2, \quad (5)$$

including only strata with $|\mathcal{B}_r| \geq 2$ and within-stratum $\text{Var}(R_T) > \varepsilon$. Mean-normalization removes RPM scale; MSE keeps a smooth gradient at $|\mathcal{B}_r|=2$, where Pearson is degenerate at ± 1 .

Role of the Beta prior on S_C . In the real regime, S_C is not measured per clip. We sample $S_C \sim \text{Beta}(199.5, 10.5)$ rescaled to $[0.92, 0.98]$ and use the sample as the conditioning token input and as the scalar argument to the kinematic targets (V_{eff}, R_T). The prior has two limited roles: it

keeps the S_C token in a plausible range and supplies weak within-stratum order for $\mathcal{L}_{\text{ratio}}$. Both physics losses are ranking objectives, not absolute slip regression; even a ± 0.03 slip perturbation changes R_T by only about 1.2 percentage points at $\theta=45^\circ$. The Beta prior is therefore a regularizer, not a supervision oracle, and real-slip claims do not rely on it as ground truth. A wider prior would strengthen $\mathcal{L}_{\text{ratio}}$ gradients but would feed real clips slip values unsupported by the apparatus; a narrower prior would be physically safer but would nearly remove the slip signal. We choose the conservative middle and evaluate true slip only in S .

Auxiliary objectives and a two-phase schedule. A latent reconstruction loss $\mathcal{L}_{\text{recon}} = \ell_1(\hat{\mathbf{x}}_0, \mathbf{x}_0)$, $\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\hat{\mathbf{x}}_0, \mathbf{x}_0)$, $\mathcal{L}_{\text{temp}}$, and small token penalties $\mathcal{L}_{\text{orth}}$, $\mathcal{L}_{\text{contr}}$ (Appendix A4.2) complement $\mathcal{L}_{\text{diff}}$. Because physics losses need coherent $\hat{\mathbf{x}}_0$, Phase 1 trains perceptual losses only, and Phase 2 ramps physics terms with $w(e)$.

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_r \mathcal{L}_{\text{recon}} + \lambda_s \mathcal{L}_{\text{SSIM}} + w(e) [\lambda_d \mathcal{L}_{\text{dyn}} + \lambda_t \mathcal{L}_{\text{temp}} + \lambda_\rho \mathcal{L}_{\text{ratio}} + \lambda_o \mathcal{L}_{\text{orth}} + \lambda_c \mathcal{L}_{\text{contr}}]. \quad (6)$$

Defaults are reported in Appendix A4.

Inference. At test time, 25-step DDIM [28] with guidance scale $s=3.0$ generates a $T=30$ -frame clip from a first frame and target (ω_p^*, S_C^*) .

5 Theoretical Insights

The losses use only scalar kinematics, but their batch form needs three checks: $\mathcal{L}_{\text{dyn}}$ must be a valid correlation objective, $\mathcal{L}_{\text{ratio}}$ must expose the weak slip axis, and D^{lat} must preserve visible-motion ordering. Full derivations are in Appendix A2.

5.1 Batch reliability of $\mathcal{L}_{\text{dyn}}$

$\mathcal{L}_{\text{dyn}}$ is the negative batch Pearson correlation between $V_{\text{eff}}(\omega_p, S_C)$ and D^{lat} (Eq. 4). It is well posed when both have nonzero batch variance. RPM-stratified sampling supplies distinct speeds, and the two-phase schedule waits until reconstructed latents have temporal motion.

Proposition 1 (Well-posed batch correlation). *For $B \geq 2$, if $s_{B, V_{\text{eff}}}^2 > 0$ and $s_{B, D^{\text{lat}}}^2 > 0$, the batch Pearson correlation in $\mathcal{L}_{\text{dyn}}$ is defined, bounded in $[-1, 1]$, locally Lipschitz in the denoiser parameters away from the zero-variance set, and differentiable almost everywhere. For $B > 3$, under a bivariate normal reference model with population correlation ρ ,*

$$\sqrt{B-3} (z_F(\hat{r}_B) - z_F(\rho)) \xrightarrow{d} \mathcal{N}(0, 1), \quad z_F(r) = \frac{1}{2} \log \frac{1+r}{1-r}. \quad (7)$$

Thus, under this asymptotic reference approximation, a Fisher-z half-width ε at confidence $1 - \delta$ corresponds to $B \approx 3 + z_{1-\delta/2}^2 / \varepsilon^2$.

At $\varepsilon=0.1$, $\delta=0.05$, $B^*=388$. Our effective batches (8 on most real runs, 16 on synthetic LOSO) therefore use $\mathcal{L}_{\text{dyn}}$ as a noisy ordering signal, not a precise population estimate. Figure 3(a,b) shows the Fisher uncertainty and nonzero denominators under stratified batching.

5.2 Why $\mathcal{L}_{\text{ratio}}$ is stratified

$R_T(S_C)$ is affine in S_C with slope $m = \sin \theta / (1 + \sin \theta) \approx 0.414$ at $\theta=45^\circ$. Under the real Beta prior, $\text{Var}(R_T | r) \approx 3.86 \times 10^{-5}$ is within an RPM.tratum, whereas V_{eff} changes by about $5\times$ from 21 to 108 RPM. A pooled slip correlation is therefore speed-dominated. Eq. (5) removes the RPM component, mean-normalizes strata, and uses MSE, so two-sample strata still have smooth gradients.

5.3 Latent and pixel motion

The physics losses use D^{lat} instead of optical flow or pixel differences. On 120 real clips (Fig. 3(c)), D^{lat} has pooled Pearson $r=0.91$ with pixel frame difference, per-RPM correlations $\{0.78, 0.83, 0.94, 0.92\}$, and low-band/high-band correlations 0.985/0.730. Thus, it captures circulation-like visible motion more strongly than self-rotation; we treat this as a scope limit, not a contact-mechanics guarantee.

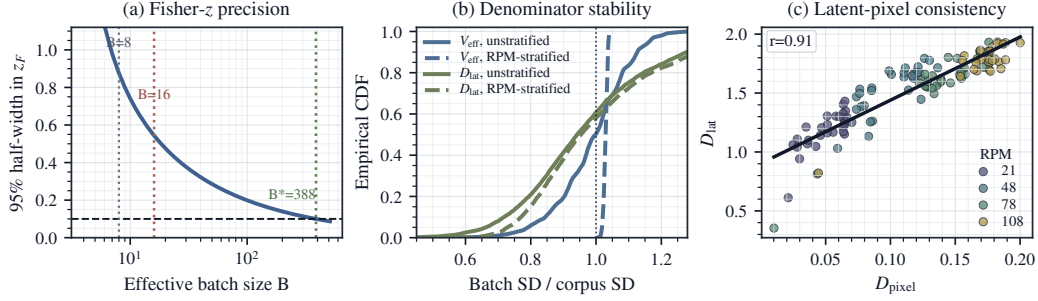


Figure 3: Foundations of the physics losses. **(a)** Fisher- z uncertainty shows that $\mathcal{L}_{\text{dyn}}$ is a noisy batch signal at our compute budget. **(b)** RPM-stratified batching keeps the $\mathcal{L}_{\text{dyn}}$ denominator away from collapse. **(c)** The latent motion scalar D^{lat} tracks pixel motion on real clips at $r=0.91$.

Scope. These arguments support scalar ordering objectives only. Generated videos are not claimed to reproduce contact mechanics, material removal, or roughness evolution; the training signal is a cross-clip kinematic prior, not a pixel-level physical law.

6 Experiments

$S_C \in [0, 1]$ denotes head-contact slip, from pure rolling (0) to pure sliding (1) [19]. Because real S_C is unmeasured, regime R evaluates realism and speed control under a narrow prior, while synthetic regime S tests slip recovery with known S_C .

6.1 Datasets and evaluation

Real regime R . Four videos at $\{21, 48, 78, 108\}$ RPM yield 604 clips of $T=30$ frames (stride 5), split 70/20/10 by source segment so overlapping windows do not cross splits. Each clip receives a persisted draw $S_C \sim \text{Beta}(199.5, 10.5)$ over $[0.92, 0.98]$ (Sec. 3); this regime grounds speed and realism but has no usable slip label.

Synthetic regime S . A renderer driven by Eqs. (1)–(2) produces matched clips at the same RPMs and known slips $\{0.20, 0.40, 0.60, 0.80\}$, supplying the wide slip variation absent from R . It is a positive-control renderer rather than a claim of photorealistic polishing: it matches the orbit geometry and timing used by the scorer but only approximates real lighting, fluid, specularities, and partial occlusions. We therefore use S to test identifiability under known labels, not to claim synthetic-to-real slip transfer. Because renderer and scorer share kinematic geometry, success on S is best read as a controlled observability test with known slip labels, not an independent physics audit; the real-domain conclusion still comes from R under the current evaluation setup.

Metrics. Visual metrics are paired PSNR, SSIM, LPIPS, FID-I3D, and FVD [11, 32]; FID-I3D uses the FVD RGB-I3D Kinetics-400 backbone, so comparisons are relative. Control uses a $4 \times 4 \times 4$ sweep per seed (scene $\times$ RPM $\times$ slip): speed rank orders $\widehat{\omega}_c$ across RPMs at fixed scene/slip, and slip rank orders $\widehat{R}_T$ across slips at fixed scene/RPM. Real sweeps use $S_C \in \{0.92, 0.94, 0.96, 0.98\}$; synthetic sweeps use the labeled S grid. Results average five seeds unless a deterministic baseline is identified.

Tracking and inversion. Scoring uses fine-tuned YOLO11n detections stitched into tracks, circulation recovery $\widehat{\omega}_c$, and Eq. (1) inversion (Fig. 4). It is independent of training: losses use latent D^{lat} and scalar $V_{\text{eff}}(\omega_p, S_C)$, while scoring uses only pixels, tracking gates (length, singularity, R^2 , cross-track MAD, sign, aggregation, plausibility), and a validation-fitted calibration c_{suite} applied before inversion. This scalar is fit once per suite on calibration clips and then fixed across variants, seeds, and test sweeps. No test labels, D^{lat} , or loss values enter scoring. Yield is the fraction of plausible sweep cells; ranks or correlations use passing cells only and are reported with yield where attrition is material. Sweeping $r_{t,\min} \in \{0.30, \dots, 0.90\}$ with an upper bound 1.35 keeps slip rank in 0.42–0.54 (mean 0.51), so rank order is not threshold-driven. Details are in Appendix A3.4. Because the same analytical kinematics define targets and inversion, the evaluation can still reward videos

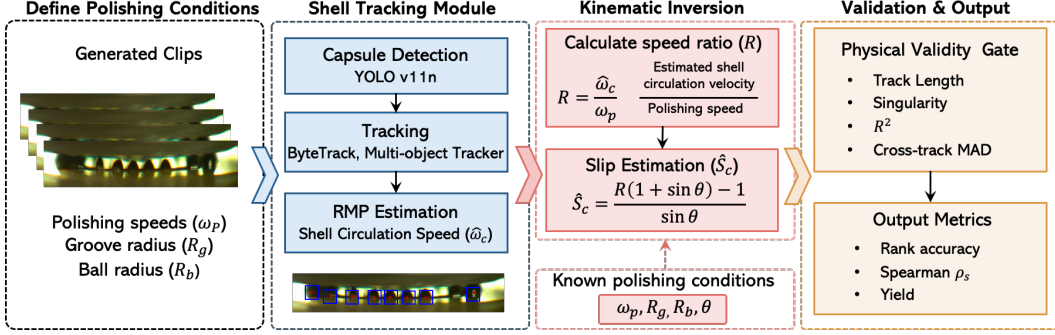


Figure 4: Slip evaluation. Generated clips pass through capsule detection, tracking, circulation recovery, kinematic inversion, and a physical validity gate before rank accuracy is computed.

Table 1: Real design study (R), five-seed mean. Visual metrics use the paired test split; speed/slip use the gated sweep. Chance slip rank is 0.5; bold marks best within group; † marks default. **Yield is the fraction of clips passing the physical validity gate.**

Group	Variant	Visual				Speed		Slip			
		PSNR↑	SSIM↑	LPIPS↓	FID-I3D↓	FVD↓	rank↑	ρ_s ↑	rank [95% CI]↑	ρ_s ↑	Yield
Loss	conditioning only	17.73	0.710	0.157	159.2	193.5	0.981	0.913	0.512 [.38,.64]	+0.024	0.956
	+ $\mathcal{L}_{\text{dyn}}$	16.58	0.688	0.185	162.5	192.7	0.975	0.912	0.508 [.39,.63]	+0.006	0.959
	+ $\mathcal{L}_{\text{ratio}}$	17.06	0.697	0.173	164.6	200.4	0.979	0.921	0.512 [.37,.65]	-0.003	0.975
	+ both ($\mathcal{L}_{\text{dyn}}, \mathcal{L}_{\text{ratio}}$) [†]	16.69	0.689	0.178	160.7	194.9	0.985	0.909	0.502 [.37,.63]	-0.033	0.959
Cond.	FiLM	16.49	0.680	0.180	171.7	204.6	0.977	0.909	0.492 [.37,.62]	-0.003	0.953
	Raw S_C (no Fourier)	15.82	0.674	0.190	161.5	192.3	0.981	0.911	0.514 [.37,.66]	-0.013	0.934
Schedule	Joint training	16.58	0.687	0.178	163.9	193.7	0.973	0.911	0.527 [.40,.65]	+0.037	0.959
	Perceptual pretrain	16.82	0.694	0.172	156.0	187.9	0.981	0.918	0.498 [.37,.63]	+0.012	0.944
	Pretrain + phys. fine tune [†]	16.75	0.690	0.176	151.6	176.2	0.975	0.906	0.500 [.37,.62]	-0.022	0.953

that produce the right centroid orbit without proving true self-spin or contact physics. We address this by keeping training and scoring signals disjoint, reporting gate yield, and treating rank scores as kinematic controllability rather than physical truth.

6.2 Real regime design study

Table 1 varies one factor at a time around the full model: cross attention, both physics losses, $K=16$, $B_{\text{eff}}=8$, and two-phase training. The full model is rerun per group, so small differences reflect five-seed variance.

Speed is robust in R : every noncollapsed variant reaches rank ≥ 0.97 . Slip stays at chance with CIs spanning 0.5, as expected for unmeasured S_C restricted to $[0.92, 0.98]$. The “conditioning only” row removes $\mathcal{L}_{\text{dyn}}, \mathcal{L}_{\text{ratio}}$ but keeps RPM and S_C tokens; its speed rank 0.981 shows that conditioning already steers speed, while physics losses give the best loss-group speed rank (0.985) and the best S FID-I3D/FVD below. Conditioning and schedule mainly tune appearance. Supplementary sweeps show stable speed rank for $B_{\text{eff}} \in \{2, 4, 8, 16\}$ and $K \in \{4, 8, 16\}$, with collapse at $K=32$.

6.3 Synthetic slip identifiability

Regime S tests slip identifiability with informative labels. Table 2 shows speed rank ≥ 0.99 and slip rank ≥ 0.735 for all settings, with CIs excluding chance.

Cross attention alone exploits the wide synthetic slip grid, so physics losses do not raise slip rank on S ; they improve distributional quality for the combined model (FID-I3D 236.4, FVD 116.0). The “conditioning only” row is not image-only: it still receives Fourier RPM and slip tokens, and clean synthetic labels make the ordinal slip axis easy for conditioning alone. The losses optimize noisier surrogates: $\mathcal{L}_{\text{dyn}}$ is speed-dominated, while $\mathcal{L}_{\text{ratio}}$ depends on D^{lat} , which is more sensitive to total visible displacement than to slip-specific circulation or self-spin. With finite batches, this can trade a few rank points for better video-distribution metrics. Learned directional flow or video features may

Table 2: Synthetic loss sweep (S), five-seed mean, plausibility gate, known $S_C \in \{0.20, 0.40, 0.60, 0.80\}$. Bold marks best per column. Yield is the fraction of clips passing the physical validity gate.

Losses	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID-I3D $\downarrow$	FVD $\downarrow$	Speed rank $\uparrow$	Slip rank [95% CI] $\uparrow$	Slip $\rho_s \uparrow$	Yield
conditioning only	21.39	0.733	0.203	243.7	126.9	0.996	0.769 [.63,.90]	+0.615	0.678
+ $\mathcal{L}_{\text{dyn}}$	21.61	0.719	0.197	246.1	127.9	0.992	0.736 [.59,.86]	+0.543	0.669
+ $\mathcal{L}_{\text{ratio}}$	21.36	0.723	0.200	245.5	126.4	0.996	0.766 [.63,.88]	+0.550	0.688
+ $\mathcal{L}_{\text{dyn}}$ + $\mathcal{L}_{\text{ratio}}$	21.58	0.734	0.201	236.4	116.0	0.996	0.735 [.59,.87]	+0.507	0.666

Table 3: Leave-one-speed-out generalization, five-seed mean, plausibility gate. R has unobserved slip; S has known $S_C \in \{0.20, 0.40, 0.60, 0.80\}$.

Held out	Type	Visual				Speed	Slip	Yield
		PSNR↑	SSIM↑	LPIPS↓	FID-I3D↓	rank [95% CI]↑	rank [95% CI]↑	
Regime R (real polishing; GT slip unobservable)								
21 RPM	boundary	11.66	0.540	0.315	147.1	0.631 [.52,.74]	0.492 [.34,.64]	0.884
48 RPM	interior	11.81	0.547	0.307	144.4	0.827 [.75,.90]	0.487 [.34,.64]	0.822
78 RPM	interior	12.78	0.581	0.279	175.0	0.875 [.81,.93]	0.588 [.44,.73]	0.922
108 RPM	boundary	12.01	0.561	0.298	169.7	0.640 [.61,.67]	0.485 [.36,.60]	0.981
Regime S (synthetic; GT slip known)								
21 RPM	boundary	20.44	0.725	0.238	228.2	0.696 [.42,.92]	0.680 [.44,.89]	0.416
48 RPM	interior	20.12	0.670	0.241	253.9	0.938 [.86,.99]	0.722 [.56,.86]	0.634
78 RPM	interior	19.75	0.660	0.250	249.5	0.721 [.62,.82]	0.742 [.62,.86]	0.709
108 RPM	boundary	20.48	0.719	0.235	233.0	0.665 [.65,.68]	0.723 [.60,.83]	0.884

help, but scalar flow magnitude is insufficient because it measures the amount of motion, not whether motion follows $R_T(S_C)$ (Appendix A3.2). Thus, slip is learnable when exposed; the real bottleneck is S_C observability.

6.4 Leave one speed out

We hold out one RPM during training and evaluate the full model on the held-out fold. Interior folds (48, 78 RPM) test interpolation; boundary folds (21, 108 RPM) test extrapolation.

Both regimes show easier speed interpolation than boundary extrapolation. Slip diverges by regime: real LOSO remains at chance, while synthetic LOSO is above chance on all folds and excludes chance on three of four.

6.5 Non-generative flow baselines

Two training-free flow-warping baselines test whether simple kinematics can replace generation. **GT-match** warps the first real frame with sinusoidal displacement matched to ground-truth flow magnitude at each RPM, making slip flat by construction. **Analytical** uses $\omega_c = \omega_p(1 + S_C \sin \theta)/(1 + \sin \theta)$ from Eq. 1, activating slip tautologically. The baselines are deterministic sanity checks on $S_C \in \{0.30, 0.50, 0.70, 0.95\}$. To avoid mixing protocols, our model row reuses the five-seed real-regime full-model result from Table 1; real slip therefore remains at chance in both tables.

Our model has higher paired fidelity than flow baselines (PSNR 16.69 vs. 13.60/12.37), but real slip rank is chance (0.502, CI spanning 0.5), matching Table 1. The analytical ceiling (0.804) is tautological and inaccessible without prescribing the true orbit. Flow baselines win FID-I3D/FVD because they warp real frames; PSNR/SSIM/LPIPS favor the generator. The real-data takeaway is speed controllability and visual fidelity, not demonstrated slip controllability.

7 Discussion

The experiments separate a positive control from an identifiability limit. Speed control is reliable: noncollapsed R variants reach rank ≥ 0.97 , and the same recipe reaches ≥ 0.99 in S . A useful envelope is $K \in \{4, 8, 16\}$, cross attention, and effective batch sizes 4–16; the schedule mainly tunes FID-I3D/FVD. Slip control depends on labels: S reaches rank ≥ 0.735 with CIs excluding

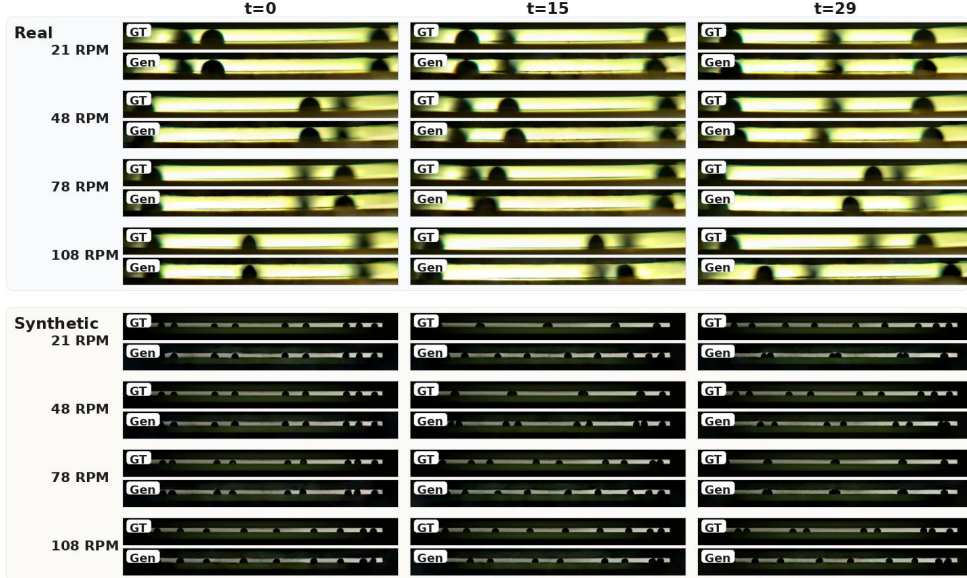


Figure 5: Representative RPM sweep for the full model. Displacement increases with RPM, consistent with the robust speed axis ranking in Table 1; slip effects are subtler and are evaluated through tracking and Lee inversion. Appendix A3.6 shows every frame of a generated clip at each speed for both regimes.

Table 4: Non-generative flow baselines and the five-seed real-regime model result. Baselines are deterministic wide-slip sanity checks; the model row is the same narrow-prior real result as Table 1, including its gate yield of 0.959.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID-I3D $\downarrow$	FVD $\downarrow$	Slip rank [95% CI] $\uparrow$
GT-match (flow)	13.60	0.679	0.248	41.8	37.9	0.484 (at chance)
Analytical (kinematic)	12.37	0.591	0.306	43.3	26.1	0.804 (tautological †)
Ours (full model, 5 seeds)	16.69	0.689	0.178	160.7	194.9	0.502 [.37, .63]

† The analytical baseline drives the warp with the same kinematic equation the scorer inverts, making its rank accuracy a tautological ceiling, not an achievable target.

chance, while R stays at chance because S_C is unmeasured and restricted to $[0.92, 0.98]$. We claim slip learnability under informative labels, not the transfer of synthetic slip control to real polishing.

Main limitations are a small number of video clips 604 from one real regime, fewer passing plausibility cells in R than in S , weaker boundary LOSO generalization, and lower held-out-speed appearance. The surrogate also stops short of being a full process-optimization tool: it predicts plausible kinematic videos, not material removal, roughness, or defect formation. Its immediate use is exploratory design-of-experiments support, sanity-checking operating sweeps, and stress-testing downstream trackers; closed-loop polishing decisions would require independent metrology linking generated kinematics to surface outcomes. Boundary LOSO failures are consistent with extrapolating both appearance and speed statistics beyond the observed range; interior folds mostly test interpolation and are more stable. This framing keeps the practical claim modest under current sensing constraints and data: the model is useful for asking whether a requested operating change produces plausible video kinematics, not for certifying a polishing recipe by itself.

Self-spin ω_b and unobservable DOF. The kinematic model predicts orbit (ω_c) and self-spin (ω_b). Our centroid tracker recovers ω_c , but ω_b appears mainly as surface texture and needs multi-angle cameras or model-based texture analysis. Since D^{lat} tracks pixel motion ($r=0.91$) but emphasizes circulation, the model is a kinematic video surrogate, not a contact-mechanics or material-removal simulator.

$K=32$ collapse. With only four training RPMs, $K=32$ introduces harmonics at $k \geq 16$ that exceed the four-point Nyquist rate, encouraging grid overfit and collapsed speed rank. $K=16$ is the recommended default.

Compute and scalability. Each run costs 12–16 GPU-hours on one RTX 4090; inference takes 35–45 s per 30-frame 496×48 clip. Rank-32 LoRA adds ≈ 22 M trainable parameters to frozen 1.5B-parameter SVD-xt. Larger slip grids or geometries need more data, but only the kinematics must be re-parameterized.

Future directions. An EM loop alternating latent slip estimation with model updates may reduce the need for per-clip slip measurement; S shows the generator can recover slip when labels exist. Learned keypoints could recover both ω_c and ω_b , and the renderer enables virtual sweeps over groove angle, capsule size, and speed before physical trials.

Broader impacts. The work targets manufacturing video surrogates for ICF capsule polishing, where controllable generation may reduce process tuning costs. It does not model implosion physics or add information beyond public polishing kinematics, but ICF has dual-use exposure and release plans that should be reviewed accordingly. The main misuse risk is treating generated clips as physical evidence beyond the scope; we separate perceptual and kinematic metrics, state that real slip is unmeasured, and require independent validation before closed-loop use.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under the contract DE-AC52-07NA27344 and by the LLNL LDRD program under Project Number 23-ERD-014.

This material is partially based on work supported while serving at the National Science Foundation. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The authors used ChatGPT (OpenAI) and Claude (Anthropic) to assist with literature review, problem formulation, coding, experiment design, analysis, interpretation of results, and manuscript preparation. The authors reviewed and edited the resulting content and take full responsibility for the methods, results, and conclusions reported in this paper.

References

- [1] H Abu-Shawareb, R Acree, P Adams, J Adams, B Addis, R Aden, P Adrian, BB Afeyan, M Aggleton, L Aghaian, et al. Achievement of target gain larger than unity in an inertial fusion experiment. *Physical review letters*, 132(6):065102, 2024.
- [2] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12467–12479, 2025.
- [3] Jan-Hendrik Bastek, WaiChing Sun, and Dennis Kochmann. Physics-informed diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [4] 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.
- [5] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6178–6189, 2025.
- [6] Minkyung Cho, Ruben Ohana, Christian Jacobsen, Adityan Jothi, Min-Hung Chen, Z Morley Mao, and Ethem Can. Tc-lora: Temporally modulated conditional lora for adaptive diffusion control. *arXiv preprint arXiv:2510.09561*, 2025.

- [7] Gunnar Farneback. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.
- [8] Ronald A Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.
- [9] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
- [10] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023.
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.
- [15] Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. *Advances in Neural Information Processing Systems*, 37:130291–130323, 2024.
- [16] Omar A Hurricane, PK Patel, R Betti, Dustin H Froula, SP Regan, Stephen A Slutz, MR Gomez, and Mary Ann Sweeney. Physics principles of inertial confinement fusion and us program overview. *Reviews of Modern Physics*, 95(2):025005, 2023.
- [17] Robert L Jackson, Hamed Ghaednia, Hyeon Lee, Amir Rostami, and Xianzhang Wang. Contact mechanics. *Tribology for scientists and engineers: from basics to advanced concepts*, pages 93–140, 2013.
- [18] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.
- [19] Rong-Tsong Lee, Yih-Chyun Hwang, and Yuang-Cherng Chiou. Lapping of ultra-precision ball surfaces. part i. concentric v-groove lapping system. *International Journal of Machine Tools and Manufacture*, 46(10):1146–1156, 2006.
- [20] Rong-Tsong Lee, Yih-Chyun Hwang, and Yuang-Cherng Chiou. Lapping of ultra-precision ball surfaces. part ii. eccentric v-groove lapping system. *International Journal of Machine Tools and Manufacture*, 46(10):1157–1169, 2006.
- [21] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024.
- [22] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.
- [23] El Moses, RN Boyd, BA Remington, CJ Keane, and R Al-Ayat. The national ignition facility: Ushering in a new age for high energy density science. *Physics of Plasmas*, 16(4), 2009.
- [24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

- [25] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [27] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.
- [28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- [29] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.
- [30] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.
- [31] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [32] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.
- [33] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [34] Chen Wang, Chuhan Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.
- [35] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.
- [36] Zexi Wu, Baolu Li, Jing Dai, Yiming Zhang, Yue Ma, Qinghe Wang, Xu Jia, and Hongming Xu. Video2lora: Unified semantic-controlled video generation via per-reference-video lora. *arXiv preprint arXiv:2603.08210*, 2026.
- [37] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlippi: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12360–12370, 2025.
- [38] Hu Ye, Jun Zhang, Sibor Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- [39] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023.
- [40] Ke Zhang, Cihan Xiao, Jiacong Xu, Yiqun Mei, and Vishal M Patel. Think before you diffuse: Infusing physical rules into video diffusion. *arXiv preprint arXiv:2505.21653*, 2025.
- [41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [43] Tianyuan Zhang, Hong-Xing Yu, Rundui Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2024.
- [44] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022.
- [45] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025.

Appendix

A1 Kinematic Equations

This supplementary section derives Eqs. (1)–(2) from first principles, following Lee et al. [19, 20]. The analysis assumes a rigid ball of radius R_b resting in a concentric V-groove of half-angle θ and groove radius R_g cut into a baseplate spinning at ω_p , with an upper lap loaded against the ball. Contact A is the inner groove wall, B the outer groove wall, and C is the head contact with the upper lap.

A1.1 Geometry

The ball contacts A and B at radial distances

$$R_i = R_g - R_b \cos \theta, \quad R_o = R_g + R_b \cos \theta, \quad (8)$$

from the baseplate axis, and C at a radial distance R_g . The ball’s spin axis tilts by β with $\tan \beta = R_b \sin \theta / R_g$; β depends only on geometry.

A1.2 No-slip at the groove walls

Let ω_c be the circulation angular speed about the baseplate axis and ω_b the self-spin. Linear contact velocities at A and B from baseplate motion must be balanced by ball-surface velocities:

$$R_i \omega_p = R_i \omega_c + R_b \omega_b \sin(\theta - \beta), \quad R_o \omega_p = R_o \omega_c + R_b \omega_b \sin(\theta + \beta). \quad (9)$$

These are the no-slip conditions at A and B .

A1.3 Slip at the head contact

At contact C ,

$$V_C = R_g \omega_{c,\text{head}} - R_g \omega_c = S_C R_g \omega_p, \quad (10)$$

where $\omega_{c,\text{head}}$ is the no-slip ball-circulation speed and $S_C \in [0, 1]$ is the dimensionless head-contact slip ratio. Combining Eq. (8)–(10) and using standard trigonometric identities yields

$$\omega_c = \omega_p \frac{1 + S_C \sin \theta}{1 + \sin \theta}, \quad \omega_b = \frac{R_g \omega_p (1 - S_C)}{R_b \cos \beta (1 + \sin \theta)}, \quad (11)$$

which, together with Eq. (10), gives the three components of Eq. (1).

A1.4 Effective sliding speed and slip-ratio target

Aggregating the three component *surface speeds*, $V_{\text{eff}} \triangleq \sqrt{(R_g \omega_c)^2 + (R_b \omega_b)^2 + V_C^2}$ is the first half of Eq. (2). Multiplying the angular velocities by their physical radii keeps all terms in m/s, so the target is invariant to the choice of length unit. Since each component speed scales linearly with ω_p for fixed S_C , V_{eff} is linear in ω_p with zero intercept and is strictly increasing in ω_p , which is the linearity used by $\mathcal{L}_{\text{dyn}}$ (Eq. 4). The circulation-to-plate ratio $R_T(S_C) = \omega_c/\omega_p = (1 + S_C \sin \theta)/(1 + \sin \theta)$ is strictly increasing in S_C for $\theta \in (0, \pi/2)$, providing the analytical ordinal target used by $\mathcal{L}_{\text{ratio}}$ (Eq. 5).

A2 Theory Derivations

This supplementary section backs Sec. 5: (i) well-posedness and asymptotic calibration of $\mathcal{L}_{\text{dyn}}$ (Sec. 5.1); (ii) the variance-decomposition argument for why $\mathcal{L}_{\text{ratio}}$ is stratified (Sec. 5.2); and (iii) the latent-pixel consistency diagnostic, with the empirical instantiation $r=0.91$ on $n=120$ clips (Sec. 5.3).

A2.1 Well-posedness and asymptotic calibration of $\mathcal{L}_{\text{dyn}}$

Setup. Write $V_i \triangleq V_{\text{eff}}(\omega_p^{(i)}, S_C^{(i)})$ and $D_i \triangleq D_i^{\text{lat}}$ (Eq. 4). The finite-batch non-degeneracy hypotheses are $s_{B,V}^2 > 0$ and $s_{B,D}^2 > 0$, where $s_{B,V}^2 \triangleq (B-1)^{-1} \sum_i (V_i - \bar{V})^2$ and analogously for D . A sufficient practical condition for $s_{B,V}^2 > 0$ is that the batch contain at least two realized speeds V_{eff} ; our stratified sampler enforces this by placing multiple physical RPM strata in each batch.

Well-posedness. The batch-Pearson correlation is

$$\hat{r}_B = \frac{\sum_{i=1}^B (V_i - \bar{V})(D_i - \bar{D})}{\sqrt{\sum_i (V_i - \bar{V})^2} \sqrt{\sum_i (D_i - \bar{D})^2}}. \quad (12)$$

The empirical variances $\frac{1}{B} \sum_i (V_i - \bar{V})^2$ and $\frac{1}{B} \sum_i (D_i - \bar{D})^2$ are strictly positive for $B \geq 2$ under the hypotheses above, so the denominator is positive and $\hat{r}_B \in [-1, 1]$ is well-defined. $\mathcal{L}_{\text{dyn}} = -\hat{r}_B$ inherits well-posedness. The numerator and denominator are locally Lipschitz functions of $\{D_i\}$ away from the zero-variance set. If the denoiser outputs are locally Lipschitz functions of the parameters, the composition is locally Lipschitz in the denoiser parameters away from the zero-variance set. Because each D_i is an ℓ_1 latent frame-difference, differentiability is almost everywhere: the only additional excluded set is where an adjacent latent pixel difference is exactly zero.

Asymptotic calibration via Fisher’s z -transform. For bivariate-normal (V, D) with correlation $\rho \in (-1, 1)$, Fisher’s z -transform $z_F(r) = \frac{1}{2} \log \frac{1+r}{1-r} = \tanh^{-1}(r)$ satisfies, classically [see e.g. 8],

$$\sqrt{B-3} (z_F(\hat{r}_B) - z_F(\rho)) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } B \rightarrow \infty. \quad (13)$$

An asymptotic $(1 - \delta)$ -CI on the z_F -scale therefore has a half-width $z_{1-\delta/2}/\sqrt{B-3}$.

Setting this asymptotic half-width to be at most ε corresponds to the reference size $B^* = \left\lceil 3 + z_{1-\delta/2}^2/\varepsilon^2 \right\rceil$. For $\varepsilon=0.1, \delta=0.05, z_{1-\delta/2}=1.96$, so $B^* = \lceil 3 + (1.96)^2/0.01 \rceil = 388$.

With only finite fourth moments, the sample covariance vector is still asymptotically normal by the multivariate CLT and $\hat{r}_B$ follows by the delta method [33], but the variance is distribution-dependent. Thus the $1/(B-3)$ formula is a Gaussian reference calculation, not an exact finite-batch interval for our four-RPM grid.

Stratified-batching corollary. If $\{\omega_p^{(i)}\}_{i=1}^B$ is constructed so each batch contains at least one sample in each of a fixed set of ω_p -strata, the empirical variance $\frac{1}{B} \sum_i (V_i - \bar{V})^2$ is bounded below by the between-strata variance of V_{eff} , a deterministic constant $\sigma_{\min}^2 > 0$. This tightens the denominator of $\hat{r}_B$ in the finite batch and prevents collapse of the correlation objective. We use this corollary only as a non-degeneracy argument for the training loss, not as a claim that the resulting batch correlation is an unbiased or lower-variance estimator of a population correlation. Our experimental practice of stratified sampling across the four speeds is consistent with this corollary. $\square$

A2.2 Variance decomposition for the stratified $\mathcal{L}_{\text{ratio}}$

Setup. A training batch $\mathcal{B}$ of size B partitions into RPM strata $\{\mathcal{B}_r\}_{r \in \mathcal{R}_B}$ with $|\mathcal{B}_r| \geq 2$. Within each stratum, samples share $\omega_p=r$ and have i.i.d. $S_C^{(i)} \sim \text{Beta}(\alpha, \beta)$ with $\mu_{S_C}=\alpha/(\alpha+\beta)$ and $\sigma_{S_C}^2 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. For $\text{Beta}(199.5, 10.5)$, $\mu_{S_C}=0.95$, $\sigma_{S_C}=0.0150$, nominal high-density range approximately $[0.92, 0.98]$.

Variance of $R_T(S_C)$ within and across strata. $R_T(S_C)$ is affine in S_C with slope $m \triangleq \sin \theta / (1 + \sin \theta) \approx 0.414$ at $\theta=45^\circ$, so

$$\text{Var}(R_T)_{\text{within}} = m^2 \sigma_{S_C}^2 \approx 3.86 \times 10^{-5}. \quad (14)$$

Because R_T depends on S_C but not directly on ω_p , pooling RPM strata does not substantially increase the variance of R_T when the same S_C prior is used in each stratum. The problem is instead that D^{lat} inherits a much larger speed-driven component: V_{eff} varies by $\sim 5\times$ from 21 to 108 RPM, whereas the within-stratum variation of R_T under the real Beta prior is only $m^2 \sigma_{S_C}^2 \approx 3.86 \times 10^{-5}$. Decomposing $D^{\text{lat}} = \alpha_v V_{\text{eff}} + \alpha_R R_T + \xi$ with mean-zero ξ ,

$$\text{Cov}_{\text{pooled}}(D^{\text{lat}}, R_T) = \alpha_R \text{Var}(R_T)_{\text{pooled}} + \alpha_v \text{Cov}(V_{\text{eff}}, R_T)_{\text{pooled}}. \quad (15)$$

When S_C is sampled independently of RPM, $\text{Cov}(V_{\text{eff}}, R_T)_{\text{pooled}}$ is small in expectation relative to the RPM-driven variance in V_{eff} , while the Pearson denominator for D^{lat} is dominated by $\alpha_v^2 \text{Var}(V_{\text{eff}})_{\text{pooled}}$. Thus the pooled correlation can underweight the weak slip-axis variation, motivating the within-RPM stratified loss.

Mean-normalized within-RPM stratification. Let $\bar{D}^{\text{lat}}_{\mathcal{B}_r} \triangleq |\mathcal{B}_r|^{-1} \sum_{i \in \mathcal{B}_r} D_i^{\text{lat}}$ and $\bar{R}_{T_{\mathcal{B}_r}}$ analogously, and define $\tilde{D}^{(i)} = D_i^{\text{lat}} / \bar{D}^{\text{lat}}_{\mathcal{B}_r}$, $\tilde{R}^{(i)} = R_T(S_C^{(i)}) / \bar{R}_{T_{\mathcal{B}_r}}$. Within a stratum V_{eff} is nearly constant because the RPM is fixed and the remaining S_C -dependence through V_C , ω_c , and ω_b is small under $\sigma_{S_C}=0.015$. So, the dominant axis of within-stratum variation in D^{lat} is whatever S_C -dependence the model produces. Mean-normalization removes the absolute scale, and $\mathcal{L}_{\text{ratio}}$ (Eq. 5) penalizes within-stratum order mismatches between latent dynamics and analytical R_T .

- *Why MSE rather than Pearson within the stratum?* For $|\mathcal{B}_r|=2$, sample Pearson is ± 1 with probability 1 and its gradient is degenerate; MSE between mean-normalized signals stays smooth.
- *Why mean-normalization rather than centering?* Both remove the across-stratum offset; mean-normalization additionally normalizes within-stratum scale, keeping $\mathcal{L}_{\text{ratio}}$ dimensionless and comparable across strata. Centering would re-introduce an RPM-dependent weight per stratum.

Filtering criterion. We exclude strata with $|\mathcal{B}_r| < 2$ or $\text{Var}_{\mathcal{B}_r}(R_T) \leq \varepsilon_{S_C}$, since these cannot contribute slip information. The stratified form is therefore a design constraint that prevents RPM-dominated variance from swamping the slip gradient at our batch size, not a separate proof of slip recovery.

A2.3 Latent-pixel consistency diagnostic

Setup. Let $\mathcal{E} : \mathbb{R}^{T \times 3 \times H \times W} \rightarrow \mathbb{R}^{T \times C \times h \times w}$ be the frozen VAE encoder. We define

$$D_i^{\text{pix}} = \frac{1}{(T-1) \cdot 3HW} \sum_{t=1}^{T-1} \|\mathbf{v}_i^{(t+1)} - \mathbf{v}_i^{(t)}\|_1, \quad (16)$$

$$D_i^{\text{lat}} = \frac{1}{(T-1)Chw} \sum_{t=1}^{T-1} \|\mathbf{z}_i^{(t+1)} - \mathbf{z}_i^{(t)}\|_1,$$

where $\mathbf{z}_i = \mathcal{E}(\mathbf{v}_i)$. Because the diffusion model is trained in latent space, the physics losses use D^{lat} rather than optical flow or pixel differences. The required contract is therefore not pixel-level physical fidelity, but preservation of the ordering of clip-level motion magnitudes.

Diagnostic criterion. We test this contract empirically by measuring the Pearson correlation between D^{pix} and D^{lat} on held-out real clips. A high positive correlation indicates that clips with larger pixel-level temporal change also tend to have larger latent temporal change, making D^{lat} a suitable

Table 5: Empirical Pearson r between D^{pix} and D^{lat} (Sec. 5.3).

Split / band	r	p -value
Overall ($n=120$, four RPMs pooled)	0.907	3.30×10^{-46}
RPM 21 (within-group)	0.778	n/a
RPM 48 (within-group)	0.827	n/a
RPM 78 (within-group)	0.939	n/a
RPM 108 (within-group)	0.920	n/a
Low band (2×2 pooled, circulation signal)	0.985	n/a
High band (residual, self-rotation signal)	0.730	n/a

scalar proxy for the ordering objective used by $\mathcal{L}_{\text{dyn}}$. This diagnostic does not prove that latent dynamics recover contact mechanics; it only verifies that the latent scalar preserves the dominant ordering of visible motion on the real clips used in this study.

Empirical instantiation. We sampled $n=120$ real polishing clips total (30 per RPM) of $T=14$ frames, encoded them through the frozen SVD-xt VAE, and measured D^{pix} and D^{lat} per clip. Table 5 reports the resulting correlations. As a reproducibility check, recomputing V_{eff} with the surface-speed norm in Eq. (2) preserves the ordering of prior diagnostic runs (Pearson $r=0.997$ with the earlier audit target), while providing the dimensionally correct physical scale used in this paper.

The overall correlation $r=0.907$ supports using D^{lat} as a clip-level motion-ordering proxy for $\mathcal{L}_{\text{dyn}}$. The lower high-band correlation $r=0.730$ also exposes a limitation: latent frame differences preserve circulation like motion more reliably than residual self-rotation. This is why Sec. 5.3 states the result as an empirical scope condition rather than as a guarantee of full contact-mechanics recovery.

A3 Additional Experimental Results

This supplementary section expands the compact results in Sec. 6. The main paper reports the primary design, synthetic, and LOSO results; here we include the effective-batch-size and Fourier-bandwidth sweeps referenced in Sec. 6.2, along with metric definitions and controlled sweep details.

A3.1 Metric definitions

I3D FVD. We use RGB-I3D features trained on Kinetics-400 [32]. Pre-processing is fixed for all variants: resize to 224×224 , sample 16 frames per clip by bilinear time interpolation from the 30-frame videos, and compute the Fréchet distance between generated and reference feature distributions.

FID and paired image metrics. FID [11] is computed with the same RGB-I3D backbone used for FVD so relative comparisons do not mix feature spaces. PSNR, SSIM, and LPIPS [42] are paired against the held-out clip associated with the same first frame.

Diagnostic flow correlation. A diagnostic Pearson correlation between RPM and mean per-clip Farneback flow magnitude [7] is computed for sanity-checking, but the headline speed and slip metrics are the gated rank accuracies recovered through the tracking pipeline (Sec. 6.1).

A3.2 Alternative motion signals for $\mathcal{L}_{\text{dyn}}/\mathcal{L}_{\text{ratio}}$

We also checked optical-flow magnitude as a candidate dynamics signal. Table 6 summarizes the diagnostic correlations and controlled slip-ranking baselines. On the same 120 real clips used for the latent-pixel diagnostic, Farneback mean flow magnitude correlates strongly with pixel L1 ($r=0.933$), latent L1 ($r=0.833$), and the corrected surface-speed V_{eff} ($r=0.905$). Thus flow is a good aggregate-motion diagnostic. It does not by itself solve slip supervision, however: a scalar flow magnitude measures how much the image moves, not whether the motion follows the slip-sensitive circulation ratio $R_T(S_C)$.

The flow-magnitude warp is intentionally slip-flat: it matches the mean flow magnitude at each RPM and therefore answers the reviewer concern that the model might only be learning speed. Its rank accuracy is at chance. The analytical warp uses the Lee kinematics directly and activates the slip

Table 6: Alternative motion-signal diagnostics for $\mathcal{L}_{\text{dyn}}$ and $\mathcal{L}_{\text{ratio}}$.

Signal / baseline	Diagnostic	Value
Farneback flow vs. D^{pix}	Pearson r on real clips	0.933
Farneback flow vs. D^{lat}	Pearson r on real clips	0.833
Farneback flow vs. corrected V_{eff}	Pearson r on real clips	0.905
Flow-magnitude warp	Controlled slip rank accuracy	0.484
Analytical kinematic warp	Controlled slip rank accuracy	0.804

axis, but it is a hand-coded image warp and has weaker paired fidelity than the five-seed diffusion model row in Table 4 (PSNR 12.37 and SSIM 0.591 versus 16.69 and 0.689). Learned flow or video embeddings such as RAFT [30] and VideoMAE [31] are plausible future signals if they retain direction and track geometry, but a scalar learned flow magnitude would likely inherit the same limitation as Farneback magnitude. Complete $\mathcal{L}_{\text{dyn}}/\mathcal{L}_{\text{ratio}}$ variants with these proxies would require decoding latents or adding a frozen video encoder inside the training loop and then calibrating a vector feature to scalar physics targets. The latent ℓ_1 signal keeps the loss cheap, differentiable almost everywhere, and aligned with the SVD latent space while the analytical V_{eff}, R_T targets provide the missing physical direction.

A3.3 Tracking and kinematic inversion pipeline

The evaluation pipeline translates generated video clips into estimated kinematic quantities ($\widehat{\omega}_c, \widehat{S}_C$) using only pixel observations, with no access to latent codes or training signals. This independence is essential: a pipeline that reused training-time quantities would not constitute an external physical test of controllability.

Step 1: Detection and tracking. A YOLO11 detector fine-tuned on polishing footage localises shell centroids in each frame. Detections are linked across frames by a Mahalanobis-IOU hybrid tracker following [44], which uses detection confidence to weight the association cost. Each track k is a sequence of frame-indexed centroids and uncertainty estimates $\{(t_n, c_{x,n}, \sigma_{x,n})\}_{n=1}^{N_k}$.

Step 2: Per-track angular velocity fitting. Because the shell orbits in a plane, its x -centroid follows a cosine projection of the orbit angle. Inverting that projection and fitting the unwrapped angle linearly in time gives an unbiased angular speed estimate without requiring the full 2-D orbit to be visible.

For track k , project the centroid onto the orbit angle:

$$u_n = \frac{c_{x,n} - x_{\text{ctr}}}{r_{\text{orb}}}, \quad \theta_n = \cos^{-1}(u_n), \quad (17)$$

where x_{ctr} and r_{orb} are the orbit centre and radius estimated from the suite geometry. The raw arccos introduces a reflection ambiguity at the orbit poles; this is resolved by a sign-of- Δx unwrapping step that gives a monotonically increasing angle series $\tilde{\theta}_n$. A weighted ordinary-least-squares fit then yields the signed angular speed:

$$\hat{\omega}^{(k)} = \arg \min_{a,b} \sum_n (\tilde{\theta}_n - at_n - b)^2. \quad (18)$$

The fit quality is assessed by $R^2 = 1 - \text{SS}_{\text{res}}/\text{SS}_{\text{tot}}$. Four per-track gates remove unreliable estimates before aggregation: (i) minimum track length $N_k \geq N_{\text{min}}$; (ii) near-singularity fraction $|\{n : |u_n| > u^*\}|/N_k \leq f^*$ (tracks that spend most of their time near the arccos poles have large projection noise); (iii) minimum angular range $\max_n \tilde{\theta}_n - \min_n \tilde{\theta}_n \geq \Delta\theta_{\text{min}}$; (iv) minimum fit quality $R^2 \geq R_{\text{min}}^2$.

Step 3: Cross-track aggregation. Even after per-track gating, a small number of spurious tracks can bias a simple mean. We therefore apply a two-stage robust aggregation [27].

First, a median absolute deviation (MAD) outlier screen removes tracks whose slope deviates from the per-clip median by more than c scaled MADs:

$$\text{MAD} = \text{median}_k (|\hat{\omega}^{(k)} - \text{median}_k \hat{\omega}^{(k)}|), \text{ keep track } k \iff |\hat{\omega}^{(k)} - \text{median}_k \hat{\omega}^{(k)}| \leq c \cdot 1.4826 \cdot \text{MAD} \quad (19)$$

Second, a weighted-median aggregation over the surviving set $\mathcal{K}$ uses weights $w_k = N_k \cdot \max(R_k^2, 0)$ to up-weight long, well-fit tracks:

$$\widehat{\omega}_c = \text{sgn}(\text{majority}) \cdot |\widehat{\omega}_{(\pi)}|, \quad \pi = \min \left\{ k : \sum_{j \leq k} w_{(j)} \geq \frac{1}{2} \sum_j w_j \right\}, \quad (20)$$

where (j) indexes tracks sorted by $|\widehat{\omega}|$ and the majority sign is determined by sign-consistency voting across $\mathcal{K}$.

Step 4: Physical validity gate. The kinematic inversion in Step 5 is only interpretable when the recovered $\widehat{\omega}_c$ lies within the physically plausible range implied by the Lee model [19] for the suite’s operating conditions. Clips whose $\widehat{\omega}_c$ falls outside this window are excluded from rank accuracy; reporting only on plausible clips prevents out-of-regime tracker noise from polluting the controllability scores.

From the Lee kinematics, the circulation ratio is

$$R_T(S_C, \theta) = \frac{1 + S_C \sin \theta}{1 + \sin \theta}, \quad \omega_c = \omega_p \cdot R_T, \quad (21)$$

giving a plausible window

$$0.5 \omega_{p,\min} R_{T,\min} \leq |\widehat{\omega}_c| \leq \kappa \omega_{p,\max} R_{T,\max}, \quad (22)$$

where $R_{T,\min} / \max$ are evaluated at the suite’s slip and RPM extremes. A clip passes the gate when Eq. (22) holds. Averaged over five seeds, clip-level gate yield ranges from 93.4% to 97.5% across the nine real-regime variants in Table 1 and from 66.6% to 68.8% across the four synthetic-regime variants in Table 2, which now report yield per row. This is distinct from the within-clip track-retention rate above (the fraction of individual tracks, not clips, kept after the length, singularity, R^2 , and cross-track MAD checks), which is lower, around 28%–32% in the real regime; an earlier draft reported this track-retention figure as gate yield, which was an error.

Step 5: Kinematic inversion. Given the aggregated $\widehat{\omega}_c$ and the known operating speed ω_p , the Lee equations are inverted to recover the circulation ratio and then the contact slip:

$$\widehat{R}_T = \frac{\widehat{\omega}_c}{\omega_p}, \quad \widehat{S}_C = \frac{\widehat{R}_T (1 + \sin \theta) - 1}{\sin \theta}. \quad (23)$$

A single scalar calibration coefficient c_{suite} , estimated on a validation split without test labels, corrects small geometric bias in $\widehat{\omega}_c$ before inversion; no slip labels enter this correction.

Step 6: Rank accuracy. A surrogate is controllable along a given axis if it can *rank* operating conditions correctly, even if absolute calibration is imperfect. Rank accuracy is therefore the natural metric: it equals 1 when generated clips are perfectly ordered by the conditioning variable, 0.5 at chance, and is insensitive to monotone rescaling of $\hat{y}$.

Formally, for a group g of clips sharing the confounding variables (i.e., same scene and slip for the speed axis; same scene and RPM for the slip axis), we compute the pairwise concordance [18]:

$$\text{RankAcc}_g = \frac{|\{(i, j) : i < j, (\text{gt}_i - \text{gt}_j)(\hat{y}_i - \hat{y}_j) > 0\}|}{|\{(i, j) : i < j, \text{gt}_i \neq \text{gt}_j\}|}. \quad (24)$$

The two evaluation axes use different ground-truth and estimate pairs:

Axis	Group by	gt	$\hat{y}$
Speed	(scene, $S_{C,\text{gt}}$)	ω_p (RPM)	$\widehat{\omega}_c$
Slip	(scene, RPM)	$R_{T,\text{gt}}$	$\widehat{R}_T$

The reported value is the bootstrap mean of $\{\text{RankAcc}_g\}$ over groups, with 95% percentile-bootstrap confidence intervals (1000 resamples).

A3.4 Controlled slip sweep and physical-validity gate

For controllability, each variant generates 64 clips per seed: four scenes, four RPM values, and four slip values ($S_C \in \{0.92, 0.94, 0.96, 0.98\}$ on R , $\{0.20, 0.40, 0.60, 0.80\}$ on S). The full tracking and

inversion procedure of Sec. A3.3 is run identically on every generated clip. A track is physically valid (plausibility gate, Eq. (22)) when $\widehat{R}_T = \widehat{\omega}_c/\omega_p \in [0.8, 1.2]$. The gate is intentionally conservative: it reduces sample size but prevents tracks outside the Lee-model regime from driving the controllability scores.

A3.5 Effective-batch-size and Fourier-bandwidth sweeps

Table 7 reports the two recipe sweeps referenced in Sec. 6.2.

Table 7: Effective-batch-size and Fourier-bandwidth sweeps on R , mean over five seeds, plausibility gate. **Bold** marks the best entry within each group. Default configuration is marked $\dagger$.

Group	Variant	Perceptual fidelity					Speed axis		Slip axis	
		PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	FVD $\downarrow$	rank $\uparrow$	$\rho_s \uparrow$	rank [95% CI] $\uparrow$	$\rho_s \uparrow$
B_{eff}	2	17.52	0.714	0.166	162.0	192.7	0.975	0.907	0.492 [.36,.62]	-0.030
	4	17.68	0.715	0.166	161.0	188.9	0.977	0.915	0.513 [.38,.64]	+0.047
	8 $\dagger$	17.10	0.700	0.173	158.0	192.6	0.967	0.908	0.492 [.36,.63]	+0.004
	16	16.38	0.684	0.181	156.9	192.4	0.975	0.916	0.479 [.35,.61]	-0.010
K	4	16.54	0.685	0.180	160.2	190.3	0.969	0.902	0.498 [.38,.62]	+0.032
	8	15.95	0.671	0.193	159.1	186.2	0.971	0.900	0.503 [.38,.63]	+0.026
	16 $\dagger$	16.70	0.690	0.174	161.9	194.4	0.971	0.918	0.517 [.38,.65]	+0.021
	32	10.88	0.439	0.356	219.0	264.5	0.770	0.624	0.441 [.24,.65]	-0.098

A3.6 Qualitative generated frames across speeds

Sec. 6 reports frame-level and video-level metrics but shows only one representative frame per condition (Fig. 5). To let a reader judge temporal coherence directly, Figs. 6 and 7 show every frame of a generated clip at each of the four RPMs, for the full model (both physics losses, cross attention, pretrain plus physics fine tune) at its final training epoch. These clips come from the model’s saved qualitative sample and validation outputs at the final epoch, at a thin top-down crop resolution (496×48 pixels for R , 480×48 for S) rather than full evaluation resolution. Frames keep their native aspect ratio; each speed block lays its 30 frames out in a 10×3 grid rather than one long row, so individual frames stay large and readable without any stretching. The synthetic clips use $S_C=0.50$, a value from training-time monitoring rather than the $\{0.20, 0.40, 0.60, 0.80\}$ evaluation grid in Sec. 6.3. Across both figures, the capsule contact point moves smoothly from frame to frame with no visible flicker or identity swap, and it advances further per frame at higher RPM, consistent with the speed axis ranking in Table 1.

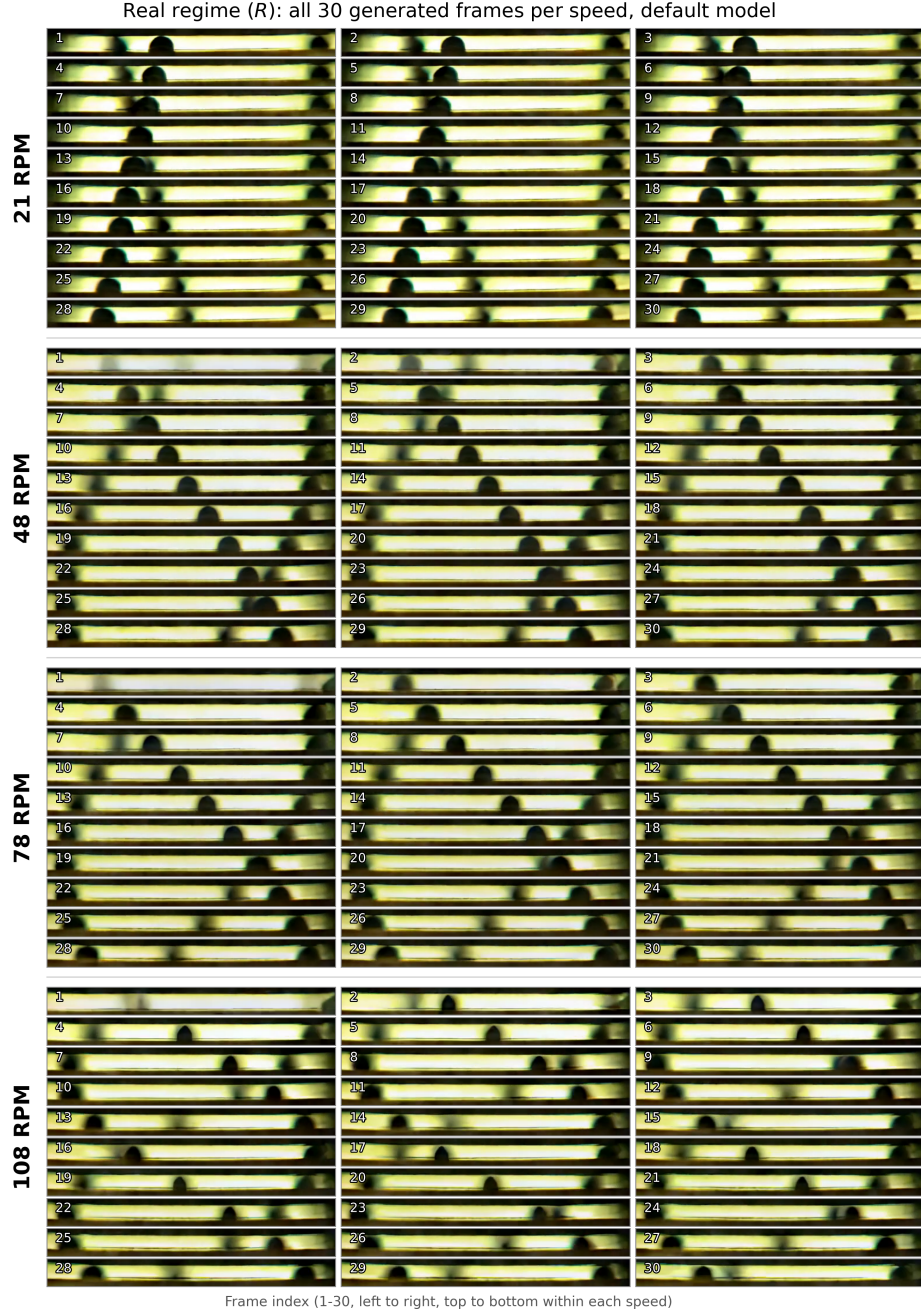


Figure 6: All 30 frames of the full model's generated clip at each real-regime (R) speed, read left to right then top to bottom within each speed block.

Synthetic regime (S): all 30 generated frames per speed, default model ($S_C = 0.50$)

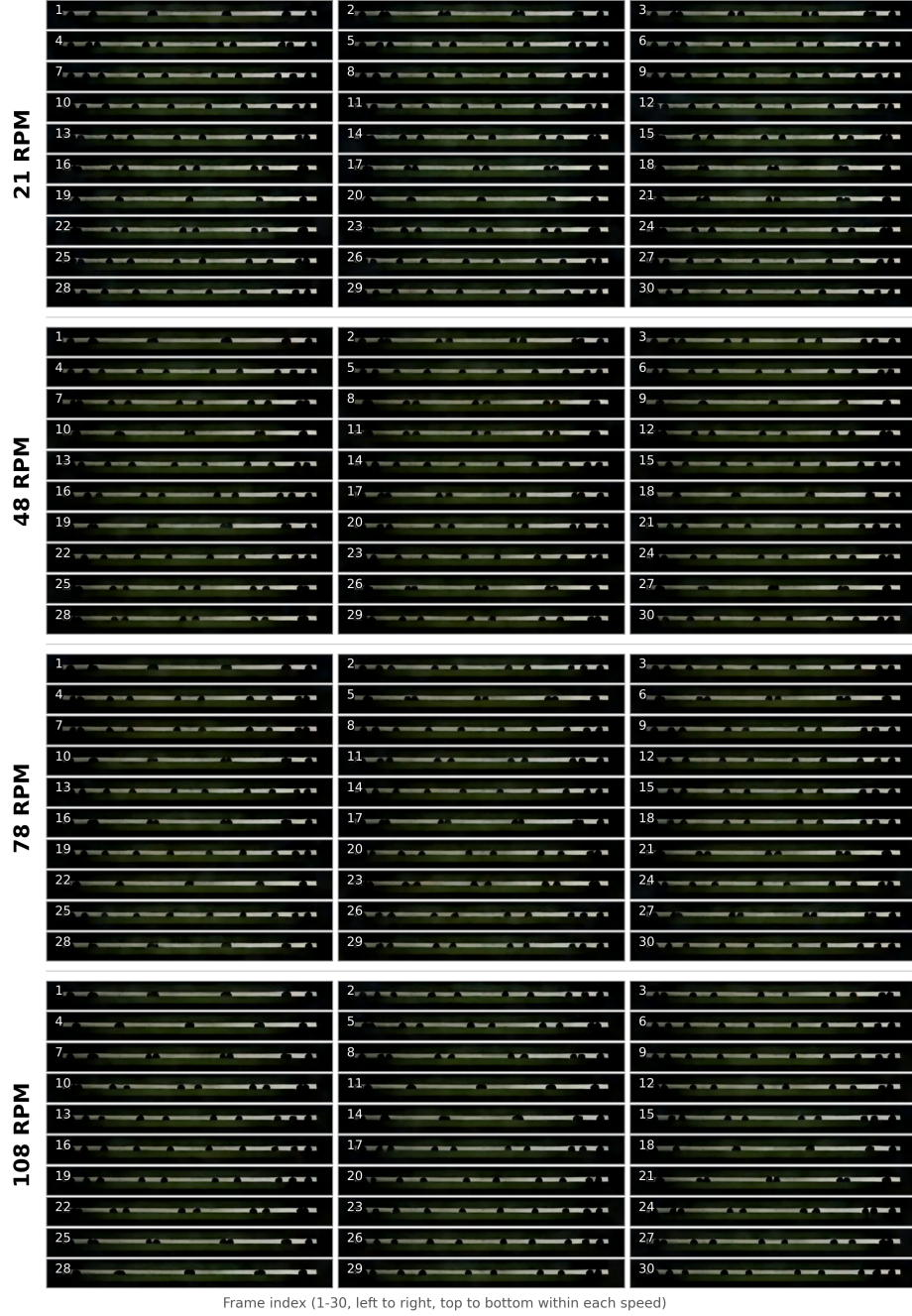


Figure 7: All 30 frames of the full model's generated clip at each synthetic-regime (S) speed, at $S_C=0.50$, read left to right then top to bottom within each speed block.

A4 Training Details

A4.1 Hyperparameters

Table 8 summarizes the training and inference hyperparameters used for the main experiments.

Table 8: Training and inference hyperparameters.

Parameter	Value
Optimization	
Optimizer	AdamW
Weight decay	0.01
LoRA learning rate	1×10^{-4}
Physics-encoder learning rate	5×10^{-4}
LR schedule	Cosine + 500-step warmup
Per-GPU batch size	2 (default real-design runs)
Gradient accumulation	4 steps (default real-design runs)
Effective batch size	8 default; audited per suite below
LoRA rank r	32
Epochs	50 (per ablation script)
GPUs	single RTX 4090 per run; multiple runs in parallel across GPUs
Conditioning	
Number of physics tokens	3 (RPM, S_C , interaction)
Fourier bands K for each scalar	16
CFG drop probability p_{cfg}	0.1
Token dim d	1024 (matches SVD-xt CLIP embedding)
Two-phase schedule	
Phase-1 epochs (perceptual only)	10
Ramp window (linear $w(e)$)	5 epochs
Loss weights (full config)	
λ_r (reconstruction)	1.0
λ_s (SSIM)	0.5
λ_d (dynamics, $\mathcal{L}_{\text{dyn}}$)	0.05
λ_t (temporal, $\mathcal{L}_{\text{temp}}$)	0.08
λ_p (ratio, $\mathcal{L}_{\text{ratio}}$)	0.05 (full); 0.0 elsewhere
λ_o (orthogonality)	0.01
λ_c (contrastive)	0.01
Inference	
Sampler	DDIM [28]
Steps	25
CFG scale s	{3.0, 5.0} (reported per-experiment)

Effective-batch-size audit. The default $B_{\text{eff}}=8$ in Table 8 refers to the real-data design-study configuration used for the main ablations and the B_{eff} sweep in Table 7. Table 9 reports the effective batch sizes used by the newer long-run suites.

Table 9: Effective batch sizes used across run families.

Run family	Per-device	Accum.	B_{eff}
Real design study default / 50 epoch	2	4	8
Real B_{eff} sweep (A5)	varies	varies	{2, 4, 8, 16}
Real full-scale 100-epoch suite	4	3	12
Real LOSO 100-epoch suite	4	3	12
Synthetic (S) loss sweep / LOSO 100-epoch suite	4	4	16

A4.2 Auxiliary loss definitions

Orthogonality regularizer. $\mathcal{L}_{\text{orth}}$ pulls the two independent physics tokens $\mathbf{T}_{\text{rpm}}, \mathbf{T}_{S_C}$ toward orthogonality so that the RPM and slip axes are encoded in separable subspaces:

$$\mathcal{L}_{\text{orth}} = |\cos(\mathbf{T}_{\text{rpm}}, \mathbf{T}_{S_C})|. \quad (25)$$

Soft contrastive regularizer. $\mathcal{L}_{\text{contr}}$ pulls RPM tokens of clips with similar V_{eff} closer in embedding space:

$$\mathcal{L}_{\text{contr}} = \frac{1}{B(B-1)} \sum_{i \neq j} w_{ij} \|\mathbf{T}_{\text{rpm}}^{(i)} - \mathbf{T}_{\text{rpm}}^{(j)}\|_2^2, \quad w_{ij} = \exp(-|V_{\text{eff}}^{(i)} - V_{\text{eff}}^{(j)}|). \quad (26)$$

A4.3 Compute budget

Table 10 summarizes the compute resources used for training and inference.

Table 10: Compute budget and trainable parameter count.

Resource	Value
Training	
Hardware	Single RTX 4090 (24 GB)
GPU-hours (50-epoch)	$\approx 6\text{--}8\text{ h}$
GPU-hours (100-epoch)	$\approx 12\text{--}16\text{ h}$
Total (all ablations)	$\approx 500\text{ GPU-hours}$
Peak GPU memory	$\approx 20\text{ GB}$
Inference	
Resolution	496×48 (raw 2040×204 , resize 0.25, aligned to $16\times$)
Frames	30
Latency per clip (DDIM 25-step)	$\approx 35\text{--}45\text{ s}$
Throughput	$\approx 80\text{--}100\text{ clips/GPU-h}$
Trainable parameters (LoRA + physics encoder)	
SVD-xt backbone (frozen)	1.5 B
LoRA (rank 32, all attn. layers)	$\approx 22\text{ M}$
Physics encoder (MLPs + interaction)	$\approx 2\text{ M}$
Total trainable	$\approx 24\text{ M}$ (1.6% of backbone)

A4.4 Latent–pixel correlation protocol

For the latent–pixel diagnostic of Sec. 5.3, we sample 120 real polishing clips total (30 at each of the four training RPMs) with $T=14$ frames, compute D^{pix} and D^{lat} per clip, and report Pearson r . Preprocessing mirrors the training path: full frame, resize factor 0.25, pixel values in $[0, 1]$, and direct VAE encoding without a $2x - 1$ pixel shift.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.

- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive” or “we were unable to find the license for the dataset we used”). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

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

Answer: [Yes]

Justification: The abstract and introduction state three contributions that match the paper’s scope: scalar-physics conditioning for SVD-xt, physics-aligned losses for speed and within-RPM slip ordering, and empirical controllability analysis. These claims are supported by the theory and experiments.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: The paper discusses concrete limitations in Section 7, including the use of 604 clips from one nominal glued-shell regime, fewer passing plausibility cells in regime R than in S , weaker boundary LOSO generalization, and reduced visual appearance quality under held-out speeds. The discussion also clarifies that real-video slip recovery is limited by unobserved per-clip slip and that the physics losses support scalar kinematic ordering rather than full contact-mechanics recovery.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or when images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed not to penalize honesty concerning limitations.

3. Theory assumptions and proofs

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

Answer: [Yes]

Justification: The paper states the assumptions and derivations in the supplement. Supplementary A1 specializes the kinematic model to derive the scalar targets V_{eff} and R_T , while Supplementary A2 covers batch nondegeneracy for $\mathcal{L}_{\text{dyn}}$, the Fisher- z asymptotic reference model, the variance argument for stratified $\mathcal{L}_{\text{ratio}}$, and the empirical latent-pixel diagnostic supporting D^{lat} as a motion-ordering proxy.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: We fully disclose the information needed to reproduce the main experimental results in Appendix A4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

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

5. Open access to data and code

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

Answer: [No]

Justification: We provide an initial version of the code in the supplementary material and plan to release all approved assets with reproduction instructions upon acceptance. Release of the real polishing videos is pending institutional approval, while synthetic videos and supporting metadata can be provided upon request.

Guidelines:

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

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

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

Answer: [\[Yes\]](#)

Justification: The paper provides the necessary experimental details in Sec. 6 and Appendix A4, including data splits, hyperparameters, optimizer settings, and evaluation procedures.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [\[Yes\]](#)

Justification: The paper reports statistical uncertainty for the main controllability metrics using confidence intervals and chance-level comparisons, including rank-accuracy intervals for speed and slip evaluation. These intervals are used to assess whether the observed ordering performance is statistically distinguishable from chance.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not include experiments.
- The authors should answer [\[Yes\]](#) if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.).
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [\[Yes\]](#)

Justification: The paper reports the compute resources needed to reproduce the experiments in Appendix A4, including GPU type, memory usage, training time, inference latency, throughput, total GPU-hours, and trainable parameter count.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The work uses process-video data from a controlled polishing setting, does not involve human subjects or personally identifiable information, and is framed for scientific modeling and process understanding rather than deployment as an autonomous safety-critical controller.

Guidelines:

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

10. Broader impacts

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

Answer: [Yes]

Justification: The paper discusses broader impacts in Section 7. It identifies positive impacts for controllable manufacturing video surrogates that may reduce process-tuning cost, while also discussing negative impacts and risks, including dual-use exposure in the ICF context, possible misuse of generated clips as physical evidence beyond their scope, and the need for independent physical validation before any closed-loop use.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [N/A]

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: All models and datasets used in our paper are all properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

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

Answer: [Yes]

Justification: The new assets introduced by the paper, including the process-video datasets, physics-loss implementation, evaluation scripts, and trained adaptation checkpoints, are documented in the supplemental material. We provide an initial version of the code in the supplementary material and plan to release all approved assets with accompanying instructions upon acceptance; release of the real polishing videos is pending institutional approval, while synthetic videos and supporting metadata can be provided upon request.

Guidelines:

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

14. Crowdsourcing and research with human subjects

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

Answer: [N/A]

Guidelines:

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

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

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

Answer: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

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

Answer: [N/A]

Guidelines:

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