Technical Report

TelePhysica: Agentic Reconstruction of Physically Interactive Scenes

Turning Any Visual Content into an Interactive Physical World

Ruiqi Shu* Xufeng Zhan* Yuanzhi Liang Haibin Huang Chi Zhang Xuelong Li

TeleAI

Abstract

Direct image-to-video diffusion models entangle pixel synthesis with implicit latent motion, precluding persistent 3D mechanical equilibrium, mid-rollout intervention, and counterfactual force control. Conversely, single-pass programmatic 3D reconstruction from an uncalibrated photograph leaves monocular depth and support clearance underconstrained, causing reconstructed scenes to interpenetrate or collapse under gravity. We present **TelePhysica**, a closed-loop framework that reconstructs an executable, multi-regime physical world from a single photograph to drive stateful 3D interaction and trajectory-grounded video synthesis. A vision-language coding agent synthesizes separable visual meshes and declarative physical models within a two-stage verification loop that pairs pixel-aligned silhouette critique with a zero-force gravitational equilibrium preflight. The verified world compiles into a unified solver coupling rigid bodies, articulated chains, co-dimensional cloth shells, and 3D tetrahedral elastodynamics, whose simulated rollouts condition pretrained video diffusion models through pixel-aligned structural streams. On a 100-scene single-image benchmark, closed-loop execution raises equilibrium-verified world recovery from 0% in single-pass synthesis to 81%. Controlled constitutive ablations show over an order-of-magnitude larger localized corner displacement and out-of-plane pleating in finite-element cloth than in matched rigid-plate proxies. Conditioning video diffusion on simulated trajectories increases mean dynamic motion-mask overlap by 2.3×–5.0× over same-scale direct image-to-video baselines and preserves motion direction under counterfactual force reversals where direct models fail.

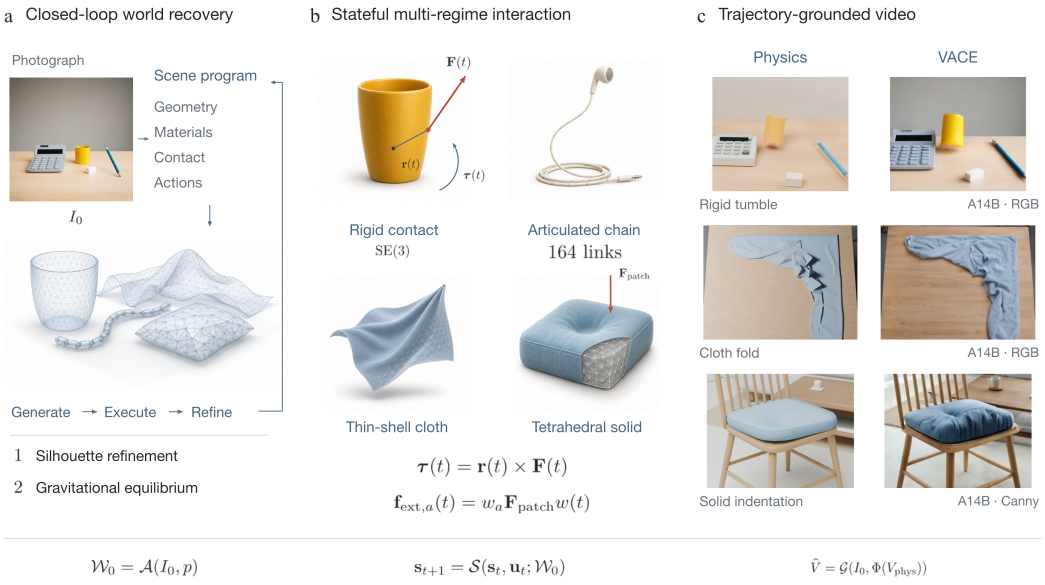


Figure 1: TelePhysica overview. (a) Closed-loop recovery from a single photograph I_0 refines a procedural scene program via Stage-1 silhouette critique and Stage-2 gravitational equilibrium checks. (b) Unified stateful mechanics spans SE(3) rigid contact, 164-link chains, thin-shell cloth, and 3D tetrahedral solids under surface wrenches $\boldsymbol{\tau}(t) = \mathbf{r}(t) \times \mathbf{F}(t)$ and distributed patch loads $\mathbf{f}_{\text{ext},a}(t) = w_a \mathbf{F}_{\text{patch}} w(t)$. (c) Phase-aligned physics and VACE frames across rigid tumbling (A14B RGB), cloth folding (A14B RGB), and solid indentation (A14B Canny).

1 Introduction

Predicting how a scene in a single photograph responds to physical interaction—tipping a ceramic mug near its rim, tugging one corner of a bath towel, or indenting an upholstered cushion—requires reasoning over 3D geometry, frictional contact, and material constitutive laws. Sequential and counterfactual interventions further demand an explicit mechanical state rather than a single synthesized clip: reversing a load vector, shifting a contact point, or applying a mid-rollout impulse alters lever-arm moments, unilateral support reactions, membrane buckling, and volumetric elastic recovery.

Existing generative and inverse-graphics paradigms face two complementary bottlenecks in interactive physical modeling. First, trajectory- and force-conditioned video generators inject text cues, 2D strokes, or predicted point tracks into diffusion denoisers [10, 40, 42, 47]. Lacking a persistent, contact-coupled 3D mechanical state, these approaches often freeze deformable surfaces, actuate the wrong object region, or fail to invert motion direction under counterfactual force reversals. Second, feed-forward 4D synthesis and open-loop inverse-graphics pipelines recover static assets or fixed neural animations [3, 5, 12, 36], while monocular projection leaves depth ordering and support clearance ambiguous along viewing rays. Placed directly under gravity without physical verification, single-pass reconstructions fail through non-manifold geometry, initial collider overlap, or support collapse.

TelePhysica resolves both bottlenecks by decoupling 3D world reconstruction, constitutive dynamics, and photorealistic video synthesis while linking them through executable verification gates (Figure 1). From a single photograph I_0 and scene description p, a vision-language coding agent synthesizes procedural visual meshes and declarative physical models inside a two-stage loop. Stage 1 aligns rendered silhouettes and interior contours with I_0 via pixel-level mask-difference critique; Stage 2 compiles candidate colliders and finite-element meshes into a dynamical solver and enforces a zero-force gravitational equilibrium gate, rejecting floating or interpenetrating layouts that appear valid in 2D projection. Verified worlds then evolve in a unified checkpointable solver spanning rigid bodies, articulated chains, thin-shell cloth, and 3D tetrahedral solids under calibrated surface wrenches, distributed patch loads, impulses, or prescribed trajectories. Finally, rendered 3D rollouts condition pretrained video generators via structural edge, shaded RGB, or motion-reference streams [1, 15, 30, 39].

Experiments across single-image reconstruction, continuum mechanics, and generative video synthesis establish three principal results:

- **Closed-loop physical world recovery.** On a 100-scene single-image benchmark, two-stage verification lifts equilibrium-verified reconstruction from 0% in single-pass open-loop synthesis (2% visual pass) to 81% in a median of 7 scene submissions, while blocking false visual-only acceptance on occluded deformable stacks.
- **Multi-regime constitutive fidelity.** Holding geometry, mass, friction, and applied load fixed, co-dimensional finite-element cloth produces 10.7× larger endpoint planar corner travel and 8–11 cm out-of-plane pleating than a matched rigid-plate proxy, while coupled 3D elastodynamics captures 5.8 cm foam indentation with sub-millimeter (0.84 mm) elastic recovery residual and soft-to-rigid momentum transfer.
- **Trajectory-grounded video synthesis.** Across the seven primary benchmark scenes, conditioning video diffusion on simulated 3D rollouts raises mean dynamic motion-mask intersection-over-union (IoU) from 0.17–0.27 in same-scale direct image-to-video baselines to 0.48–0.87, reduces mean centroid trajectory error by 3.9×–20.2×, and maintains $> +0.90$ directional cosine alignment under counterfactual force reversals where direct baselines yield negative correlation.

2 Related Work

2.1 Programmatic and Agentic 3D Scene Synthesis

Representing 3D environments as executable graphics programs provides interpretable hierarchy, separable object instances, and parametric editability. Language-guided layout and procedural modeling systems—including LayoutGPT [8], 3D-GPT [36], SceneCraft [12], and Holodeck [45]—translate text specifications into spatial arrangements or Blender scripts, while BlenderAlchemy [14] and Thinking-in-Blender [11] close a visual render loop to refine materials and geometry. Programmatic world representations also support symbolic physical reasoning in Code-as-Worlds [41] and VisPhyWorld [22], and 4DCodeBench [35] evaluates coding agents on reconstructing observed dynamic trajectories from monocular videos, complementing systems-level advances in distributed model collaboration [2, 34] and static-scene delivery [48]. Rather than fitting a single observed trajectory from multi-frame video, TelePhysica reconstructs an equilibrium-verified physical world from a single static photograph to support diverse post-reconstruction force interventions, mid-rollout continuation, and counterfactual queries.

2.2 Interactive Physical World Modeling Across Material Regimes

Coupling 3D scene representations with multi-body and continuum solvers [6, 29, 38] grounds visual synthesis in Newtonian mechanics. Feed-forward 4D generation and score-distillation pipelines [3, 5] synthesize plausible object motion but bake deformation into neural fields without explicit contact colliders. Differentiable continuum digital twins—including PAC-NeRF [20], PhysGaussian [44], PhysDreamer [51], Spring-Gaus [53], PIE-NeRF [9], PhysMotion [37], and Phys4DGen [25]—embed material-point, spring-mass, finite-element, or meshless dynamics into radiance fields or 3D Gaussian splats across multi-view and single-object settings. For single-view scene interaction, PhysGen [27] models planar rigid dynamics in image space, whereas PhysGen3D [4], WonderPlay [21], RealWonder [28], PerpetualWonder [49], and Fluid-Gen-Zero [13] lift single photographs into 3D meshes, Gaussian particles, or fluid emitters for action-conditioned rollouts. TelePhysica complements particle-based lifting by compiling agent-synthesized meshes into a unified constraint and finite-element solver that couples six-degree-of-freedom rigid bodies, high-link articulated chains, thin-shell cloth, and 3D tetrahedral solids behind a deterministic checkpoint interface.

2.3 Trajectory-Conditioned and Aligned Video Generation

Motion control in video diffusion transformers [16, 39, 46] spans latent trajectory injection and dense structural conditioning. VAST 1.0 [50] decouples storyboard layout planning from pixel generation, whereas DragNUWA [47], MotionCtrl [42], DragAnything [43], Tora [52], Force Prompting [10], and PhysCtrl [40] condition diffusion denoisers on 2D strokes, local force cues, or generative 3D point trajectories. Without an executable multi-body contact solver, however, out-of-plane rotations, multi-object collisions, and layered cloth disocclusions remain susceptible to silhouette drift and directional ambiguity. Structural video adapters such as VACE [1, 15] inject frame-aligned contour or RGB tensors alongside reference frames, while MiniMax-H3 Ref2VA [30] generates joint audio-video streams from paired appearance and motion references, and complementary post-training frameworks align generative models via reward optimization, step-level credit assignment, and trajectory shaping [17–19, 23, 24, 26, 32]. TelePhysica supplies solver-verified 3D rollouts as conditioning streams for pretrained generators and characterizes how model scale and control density govern trajectory fidelity versus material realism.

3 Method

TelePhysica maps an input photograph $I_0 \in \mathbb{R}^{H\times W\times 3}$ (height H, width W) and scene description p to an executable physical world $\mathcal{W}_0$ and trajectory-grounded video rollouts $\widehat{V}$ through three stages (Figure 1): closed-loop inverse graphics with gravitational equilibrium verification (§3.1), unified multi-regime elastodynamics under calibrated perturbations (§3.2), and deformable surface skinning coupled with structural video diffusion

(§3.3). All stages share a right-handed world frame with $+Z$ opposing gravity $\mathbf{g} = (0, 0, -9.81)^\top$ m/s^2, SI units (m, kg, s, N, Pa), and scalar-first unit quaternions $\mathbf{q}_i^{\text{rot}} = (q_w, q_x, q_y, q_z) \in S^3$ with rotation matrices $\mathbf{R}(\mathbf{q}_i^{\text{rot}}) \in \text{SO}(3)$ (Appendix A).

3.1 Closed-Loop Inverse Graphics and Physical Equilibrium Verification

Reconstructing an interactive 3D world from an uncalibrated view I_0 requires recovering camera calibration $\mathcal{M}$, separable visual meshes $O = \{o_i\}_{i=1}^N$ across N scene objects, collision primitives and finite-element rest topologies C (carrying masses, constitutive parameters, contact margins, and mobility roles $\rho_i \in \{\text{dynamic}, \text{kinematic}, \text{static}, \text{visual-only}\}$), and a statically stable initial mechanical state $\mathbf{s}_0$:

$$\mathcal{W}_0 = \big(O, C, \mathcal{M}, \mathbf{s}_0\big) = \mathcal{A}(I_0, p). \tag{1}$$

Because perspective projection collapses depth along optical rays, an object floating several centimeters above a tabletop projects to the identical 2D contour as a contacting object when scaled proportionally along its viewing ray. Open-loop single-pass synthesis therefore produces layouts that either misalign in image space or interpenetrate, topple, and drift under gravity.

To resolve projective and mechanical ambiguity jointly, the reconstruction agent $\mathcal{A}$ iterates through a two-stage verification loop inside a headless rendering and physics sandbox. The agent first defines a semantic object inventory with persistent identifiers and mobility roles, keeping movable foreground objects dynamic while locking supporting tables, floors, and walls as static colliders. In Stage 1 (*visual mask-difference critique*), the agent synthesizes a procedural 3D script specifying manifold visual meshes, physically based materials, lighting, and camera parameters. Rendered per-object silhouettes form an aggregate foreground mask M_{rend}, which is evaluated against the reference mask M_{ref} of I_0 via foreground IoU, a 72-bin radial centroid-to-contour profile, and Sobel interior edge overlap. A three-panel diagnostic overlay isolating reference-only deficits and render-only excess guides iterative updates to focal length, object dimensions, and 3D poses.

Visual alignment alone does not guarantee static support. In Stage 2 (*zero-force physical equilibrium preflight*), the candidate world compiles into the dynamical solver with zero external actuation ($\mathbf{u}_t = \mathbf{0}$) and integrates under gravity $\mathbf{g}$ over $T_{\text{eq}} = 1.0$ s. For each dynamic object i, let $d_i(t)$ and $v_i(t)$ denote center-of-mass displacement $\|\mathbf{x}_i(t) - \mathbf{x}_i(0)\|_2$ and linear speed $\|\dot{\mathbf{x}}_i(t)\|_2$ for rigid bodies, or maximum nodal displacement $\max_{a \in \mathcal{V}_i} \|\mathbf{x}_{i,a}(t) - \mathbf{x}_{i,a}(0)\|_2$ and maximum nodal speed $\max_{a \in \mathcal{V}_i} \|\dot{\mathbf{x}}_{i,a}(t)\|_2$ for deformable bodies. Acceptance requires bounded pairwise initial contact penetration $\delta_{ij}(0)$, scale-calibrated translational or nodal rest, bounded terminal speeds, and solver health across every integration step:

$$\delta_{ij}(0) \le \varepsilon_{c,ij}, \qquad \max_{t \in [0, T_{\text{eq}}]} d_i(t) \le \varepsilon_{\text{eq},i}, \qquad v_i(T_{\text{eq}}) \le \varepsilon_{v,i}, \qquad \text{finite}(\mathbf{s}_t) \ \wedge \ N_{\text{warn}} = 0, \tag{2}$$

alongside rigid-body orientation drift $\Delta\theta_i \le 5^\circ$ and terminal angular speed $\|\boldsymbol{\omega}_i(T_{\text{eq}})\|_2 \le 0.1$ rad/s, where $\varepsilon_{c,ij}$, $\varepsilon_{\text{eq},i}$, and $\varepsilon_{v,i}$ denote feature-scaled tolerances (Table 4). When Stage 2 detects collider overlap, role mismatch, or gravitational drop, structured contact and drift traces feed back to the agent for geometric revision; on multi-object deformable stacks with occluded support clearances, support repair and damped quasi-static settling relax residual contact gaps before re-verifying Stage 1 and Equation (2) (with extended 3.0–4.0 s zero-input stability checks reported in Table 2).

3.2 Unified Multi-Regime Elastodynamics and Surface Perturbations

Once verified, $\mathcal{W}_0$ maintains an explicit generalized state $\mathbf{s}_t = (\mathbf{q}_t, \mathbf{v}_t)$ with configuration coordinates $\mathbf{q}_t$, generalized velocities $\mathbf{v}_t$ ($\dot{\mathbf{q}}_t = \mathbf{N}(\mathbf{q}_t)\mathbf{v}_t$), and discrete timestep Δt via $\mathbf{s}_{t+1} = \mathcal{S}(\mathbf{s}_t, \mathbf{u}_t; \mathcal{W}_0)$. Our declarative schema compiles rigid, articulated, thin-shell, and volumetric deformable objects into a shared contact-coupled formulation [29, 38].

Multi-Regime Constitutive Formulation. Rigid bodies evolve on SE(3) with six velocity coordinates per free body and convex primitive or decomposed colliders subject to elliptic frictional contact cones. Flexible 1D structures such as coiled cables are modeled as capsule chains linked by spherical joints that enforce segment inextensibility while transmitting axial tension and surface friction. Continuum deformables are discretized on a rest finite-element mesh $\mathcal{G}_{\text{mesh}} = (\mathcal{V}, \mathcal{E})$ with $n_v = |\mathcal{V}|$ nodal positions $\mathbf{X}(t) = [\mathbf{x}_1(t), \dots, \mathbf{x}_{n_v}(t)]^\top \in \mathbb{R}^{n_v \times 3}$ and lumped masses $m_a > 0$. In generalized velocity coordinates $\mathbf{v}$, the coupled equations of motion take the form:

$$\mathbf{M}(\mathbf{q})\dot{\mathbf{v}} + \mathbf{c}(\mathbf{q}, \mathbf{v}) + \mathbf{f}_{\text{int}}(\mathbf{q}, \mathbf{v}) = \mathbf{f}_{\text{grav}}(\mathbf{q}) + \mathbf{J}_c(\mathbf{q})^\top \boldsymbol{\lambda}_c + \mathbf{f}_{\text{ext}}(t), \tag{3}$$

where $\mathbf{M}(\mathbf{q})$ combines rigid spatial inertias and deformable nodal masses, $\mathbf{c}(\mathbf{q}, \mathbf{v})$ collects Coriolis, centrifugal, and gyroscopic bias terms, $\mathbf{f}_{\text{int}}(\mathbf{q}, \mathbf{v})$ denotes the elastic-energy gradient and viscous damping resistance (corresponding to passive restoring force $-\mathbf{f}_{\text{int}}$), and $\mathbf{J}_c(\mathbf{q})^\top \boldsymbol{\lambda}_c$ resolves unilateral normal and Coulomb frictional contacts across rigid–rigid, deformable–rigid, and deformable–deformable pairs with contact margin radius r_c. Triangle meshes ($e \in \mathcal{E}$, $|e| = 3$, $d = 2$) model woven and knitted textiles through coupled in-plane membrane stretching and cross-edge bending elasticity, complemented by compliant XPBD shell projection in Appendix A. Tetrahedral meshes ($e \in \mathcal{E}$, $|e| = 4$, $d = 3$) model compressible foams and cushions via three-dimensional continuum elasticity parameterized by Young's modulus $E > 0$, Poisson's ratio $\nu \in [0, 0.5)$, and viscous damping η. Implicit-in-velocity integration advances coupled rigid and deformable states jointly with frictional contact tractions.

Calibrated Perturbations, State Continuation, and Counterfactuals. External actions $\mathbf{u}_t$ enter Equation (3) either as localized surface and patch loads or as impulsive and kinematic boundary conditions. For a 2D stroke drawn on I_0 from $\mathbf{u}_{\text{start}}$ to $\mathbf{u}_{\text{end}} \in [0, 1]^2$ with dimensionless screen length $\ell = \|\mathbf{u}_{\text{end}} - \mathbf{u}_{\text{start}}\|_2$, camera ray-casting locates the 3D surface anchor $\mathbf{x}_a(t_0)$ on rigid body o_k (center of mass $\mathbf{x}_k(t)$) and projects the stroke onto the camera-parallel plane along unit direction $\hat{\mathbf{d}} \in \mathbb{R}^3$. Weight-proportional scaling with $g = 9.81\ \text{m/s}^2$ and dimensionless gain $\alpha > 0$ yields a world-frame force and center-of-mass torque over temporal activation window $w(t) \in [0, 1]$:

$$\mathbf{F}(t) = \alpha\, \ell\, m_k\, g\, w(t)\, \hat{\mathbf{d}}, \qquad \boldsymbol{\tau}(t) = \big(\mathbf{x}_a(t) - \mathbf{x}_k(t)\big) \times \mathbf{F}(t), \tag{4}$$

where lever arm $\mathbf{r}(t) = \mathbf{x}_a(t) - \mathbf{x}_k(t)$ generates off-center tipping, pitching, and yawing moments around $\mathbf{x}_k(t)$. On deformable shells and 3D solids, manipulation acts as a distributed patch load with resultant force $\mathbf{F}_{\text{patch}} \in \mathbb{R}^3$ (in N) distributed over node subset $\mathcal{P} \subset \mathcal{V}$ via nonnegative spatial weights $w_a \geq 0$ ($\sum_{a \in \mathcal{P}} w_a = 1$) and $\sin^2$ temporal window $w(t) = \sin^2\big(\pi(t - t_{\text{on}})/(t_{\text{off}} - t_{\text{on}})\big)$ on $t \in [t_{\text{on}}, t_{\text{off}}]$:

$$\mathbf{f}_{\text{ext},a}(t) = w_a\, \mathbf{F}_{\text{patch}}\, w(t), \qquad a \in \mathcal{P}, \qquad \sum_{a \in \mathcal{P}} w_a = 1, \tag{5}$$

distributing the applied load across the selected surface patch rather than concentrating it at a single node. Instantaneous impulses apply an unconstrained velocity increment $\Delta\mathbf{v} = \mathbf{M}(\mathbf{q})^{-1}\mathbf{J}_{\text{gen}}$ at event time t_b prior to contact-constraint resolution, whereas prescribed boundary trajectories drive designated attachment links along quintic smoothstep profiles $s(\zeta) = 10\zeta^3 - 15\zeta^4 + 6\zeta^5$ for normalized phase $\zeta \in [0, 1]$. Complete solver checkpoints $\boldsymbol{\xi}_t$ serialize generalized coordinates $\mathbf{q}_t$, velocities $\mathbf{v}_t$, warm-start contact states, and active perturbation schedules, supporting reproducible trajectory replay within $\leq 10^{-12}$ state tolerance, live mid-rollout continuation at intermediate timestamp t_b, and counterfactual branching from identical initial states $\mathbf{s}_0$.

3.3 Deformable Surface Skinning and Trajectory-Conditioned Video Diffusion

To render physics rollouts $V_{\text{phys}} = \mathcal{R}(\mathbf{s}_{0:T}; \mathcal{W}_0)$ without an initial discontinuity between visual and physical poses, each rigid visual mesh preserves its initial body-to-visual offset $\mathbf{T}_{\text{rel},i} = \mathbf{T}_{\text{body},i}(0)^{-1}\mathbf{T}_{\text{vis},i}(0) \in \text{SE}(3)$ and updates as $\mathbf{T}_{\text{vis},i}(t) = \mathbf{T}_{\text{body},i}(t)\mathbf{T}_{\text{rel},i}$. Fine deformable visual meshes resolving hemmed borders, knitted ribs, or upholstery seams beyond the finite-element grid deform via precompiled linear-blend and normal-offset

skinning constructed to reproduce each visual vertex exactly at $t = 0$. Each visual vertex $\mathbf{p}_r$ binds to K simulation nodes $\{i_{r,j}\}_{j=1}^{K}$ with partition-of-unity weights $\sum_{j=1}^{K} w_{r,j} = 1$—spanning direct nodal ($K = 1$), shell barycentric ($K = 3$), and tetrahedral or trilinear cell ($K \in \{4, 8\}$) bindings—plus a signed thickness offset d_r along its surface carrier triangle (a_r, b_r, c_r) ($d_r = 0$ for volumetric interior bindings):

$$\mathbf{p}_r(t) = \sum_{j=1}^{K} w_{r,j}\, \mathbf{x}_{i_{r,j}}(t) + d_r\, \hat{\mathbf{n}}_r(t), \qquad \hat{\mathbf{n}}_r(t) = \frac{\big(\mathbf{x}_{b_r}(t) - \mathbf{x}_{a_r}(t)\big) \times \big(\mathbf{x}_{c_r}(t) - \mathbf{x}_{a_r}(t)\big)}{\big\|\big(\mathbf{x}_{b_r}(t) - \mathbf{x}_{a_r}(t)\big) \times \big(\mathbf{x}_{c_r}(t) - \mathbf{x}_{a_r}(t)\big)\big\|_2}. \tag{6}$$

A pretrained video generator $\mathcal{G}_{\text{vid}}$ then synthesizes $\widehat{V} = \mathcal{G}_{\text{vid}}\big(I_0, \Phi(V_{\text{phys}}), p_{\text{vid}}\big)$ anchored to I_0. Within structural diffusion adapters spanning Wan2.1-VACE 1.3B and Mixture-of-Experts Wan2.2-VACE-Fun A14B (14B active parameters) [1, 15, 39], we study two pixel-aligned control operators Φ: sparse Canny contours $\Phi_{\text{Canny}}(V_{\text{phys}})$, which constrain silhouettes while leaving interior texture open to reference-guided synthesis, and dense shaded renders $\Phi_{\text{RGB}}(V_{\text{phys}})$, which provide depth-ordered occlusion and surface orientation cues. On rigid scenes, we additionally evaluate dual-reference audio-video synthesis with MiniMax-H3 Ref2VA [30], where $\Phi_{\text{Ref}}(V_{\text{phys}}) = V_{\text{phys}}$, alongside image-and-text-conditioned direct image-to-video baselines prompted from I_0 without simulated trajectories.

4 Experiments

Our experiments evaluate three questions: how two-stage verification-guided refinement governs single-image 3D world recovery (§4.1), how explicit multi-regime continuum mechanics compares against rigid-body approximations (§4.2), and how simulated 3D rollouts govern trajectory fidelity and appearance preservation in video diffusion (§4.3).

4.1 Closed-Loop Agentic Reconstruction vs. Single-Pass Open-Loop Synthesis

We audit complete agent trajectories across our 100-scene single-image tabletop benchmark—covering household vessels, layered stationery, produce, lamps, footwear, and containers—together with three single-image deformable and hybrid reconstructions: a microfiber bath towel, an upholstered seat cushion on a wooden spindle-back chair, and two stacked kitchen sponges beside a detergent bottle. Table 1 and Figure 2 compare single-pass open-loop synthesis, evaluated on the agent's initial scene submission, against full two-stage closed-loop execution.

On the 100-scene benchmark, single-pass open-loop synthesis passes visual alignment on 2% of inputs and achieves 0% combined visual-and-physical verification. Two sequential failure barriers account for this result (Figure 2b). First, 98 of 100 initial scene scripts fail Stage 1 through non-manifold or misbound procedural meshes (68%), script or renderer exceptions (26%), or silhouette misalignment (4%). Second, among the 95 scenes that cross the Stage-1 visual threshold during iterative critique to enter a first Stage-2 physics preflight, 84 (88.4%) fail that initial check due to support-plane interpenetration or static/dynamic role errors (45/95), invalid quaternion or inertia specifications (36/95), or gravitational settling drift (3/95). Returning structured execution diagnostics resolves both barriers: while collider repairs on 10 borderline scenes shift final silhouettes slightly below the Stage-1 contour threshold, closed-loop execution recovers 85% final visually aligned scenes, 81% joint Stage-1 and Stage-2 equilibrium-verified world packages in a median of 7 scene submissions and 17 tool calls, and 49% verified 5.0 s interactive surface-force rollouts.

The same two-stage verification governs single-image deformable and hybrid finite-element reconstruction (Figure 2a and Figure 3). Across all three scenes, closed-loop visual critique raises foreground IoU from 0.24–0.37 to 0.52–0.70 (0.60–0.66 composite visual score) with zero initial mesh interpenetration. Both the 121-node bath towel and the 243-node tetrahedral seat cushion satisfy the Stage-2 equilibrium gate autonomously within a single agent run, maintaining < 1 mm peak nodal drift over extended 4.0 s zero-input verification. In the three-object sponge-and-bottle scene, Stage 1 reaches clean 2D alignment (IoU 0.55,



Figure 2: Two-stage closed-loop verification versus single-pass open-loop synthesis. (a) Single-image reconstruction traces on the three-object sponge-and-bottle scene and an upholstered chair cushion, showing input photographs, initial and refined renders, normalized foreground mask-difference diagnostics (red: reference-only deficit; cyan: render-only excess), and Stage-2 zero-force peak nodal drift over the stated verification horizons. On the sponges, visual refinement reaches IoU = 0.55, whereas the 1.0 s Stage-2 preflight detects a 68.3 mm nodal drop across an occluded vertical support gap on the tilted upper sponge (0.661 mm over 4.0 s after support repair and damped contact relaxation). The chair cushion passes autonomously with 0.979 mm peak nodal drift over 4.0 s. (b) Controlled 100-scene tabletop audit comparing first-submission open-loop against closed-loop execution across visual alignment (2% vs. 85%), equilibrium verification (0% vs. 81%), and full 5.0 s force rollouts (0% vs. 49%), alongside initial Stage-1 (98/100) and first Stage-2 preflight (84/95) rejection breakdowns.

silhouette structure 0.66) with zero initial overlap, yet monocular occlusion hides a vertical gap beneath the tilted upper sponge. The 1.0 s Stage-2 preflight catches the resulting 68.3 mm peak nodal drop to prevent false visual-only acceptance, and support repair with damped contact relaxation brings both sponges and the bottle within equilibrium tolerances (0.66 mm peak nodal drift and $\leq$ 0.57 mm CoM drift over 4.0 s).

4.2 Multi-Regime Stateful Mechanics, Counterfactuals, and Constitutive Ablation

Figure 4 and Table 2 analyze how constitutive laws govern contact isolation, deformability, and counterfactual interventions across rigid, articulated, thin-shell, and volumetric hybrid worlds.

Controlled Constitutive Proxy Ablation on Thin-Shell Cloth. To test whether thin flexible objects can be approximated by rigid collision proxies, we hold visual geometry, total mass (0.23 kg), tabletop Coulomb friction ($\mu = 0.80$), support geometry, and the corner load pulse ($[2.6, 2.6, 0]$ N over $t \in [0.20, 0.36]$ s) fixed on the microfiber bath towel while replacing the 121-node cloth shell with a matched rigid plate (Figure 4a). Enforcing rigid-body kinematics across the 0.76×0.80 m footprint requires overcoming normal friction across the entire sheet simultaneously, limiting motion to a 7.2 cm planar slab shift (5.1 cm per axis) with zero out-of-plane buckling. By contrast, the co-dimensional cloth shell decouples local membrane compliance from distant frictional anchorage: the actuated corner travels 10.7× farther in plane (77.2 cm endpoint planar

Table 1: Controlled evaluation of first-submission open-loop versus two-stage closed-loop reconstruction. Audited across the 100-scene single-image rigid tabletop benchmark and three single-image deformable and hybrid finite-element reconstructions.

Benchmark & Verification Metric	Open-Loop (Pass 1)	Closed-Loop	Dominant Failure Modes Diagnosed & Repaired
100-Scene Single-Image Tabletop Benchmark (N = 100)			
Stage-1 visual alignment pass	2/100 (2.0%)	85/100 (85.0%)	Pass-1 rejections: 68/100 mesh topology or instance-binding errors, 26/100 script/render exceptions, 4/100 silhouette mismatches.
Stage-1 + Stage-2 equilibrium verified	0/100 (0.0%)	81/100 (81.0%)	First physics preflight (95 scenes reaching Stage-1 during loop): 45/95 collider overlap or role errors, 36/95 pose/inertia metadata errors, 3/95 zero-force drift (11/95 pass on first check).
Full 5.0 s surface-force rollout verified	0/100 (0.0%)	49/100 (49.0%)	The 81 equilibrium-verified packages converge in median 7 scene submissions (mean 9.21) and median 17 tool calls (mean 18.72).
Single-Image Deformable & Hybrid Finite-Element Reconstructions (N = 3: Bath Towel, Seat Cushion & Chair, Kitchen Sponges & Bottle)			
Stage-1 visual + contact-clearance pass	0/3 (0.0%)	3/3 (100.0%)	Iterative camera/pose alignment reaches silhouette IoU 0.52–0.70 with zero initial mesh interpenetration across all three scenes.
Autonomous < 1 mm nodal drift pass	0/3 (0.0%)	2/3 (66.7%)	Bath towel (0.48 mm peak node drift) and chair cushion (0.98 mm peak node drift) pass autonomously; Stage-2 catches a 68.3 mm nodal drop across an occluded support gap on the tilted upper sponge prior to damped relaxation.

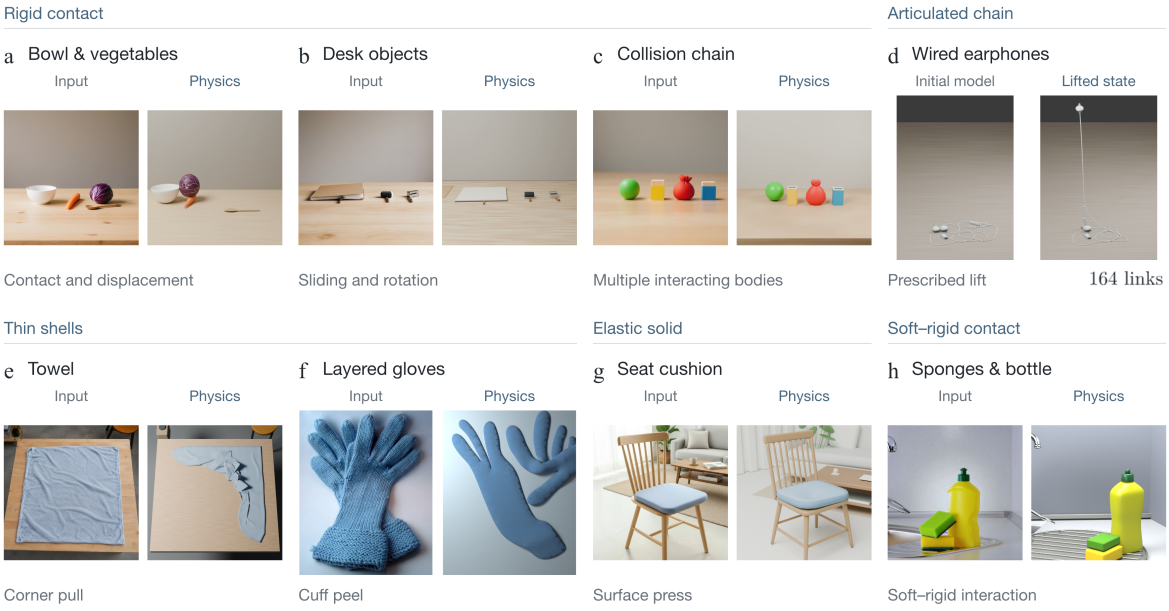


Figure 3: Single-image 3D world reconstruction and simulated interactions across constitutive regimes. (a–c) Multi-object rigid contact, sliding, and collision cascades pairing input photographs with simulated post-perturbation states. (d) A 164-link articulated earphone cable in its initial resting model state and after a prescribed 300 mm earbud lift. (e,f) Co-dimensional thin-shell towel corner folding and layered knitted glove cuff peeling. (g,h) Three-dimensional tetrahedral seat cushion indentation and soft-to-rigid sponge-to-bottle contact transfer.

displacement; 78.7 cm temporal 3D peak) and buckles into an 80.4 mm nodal pleat (108 mm skinned surface crest) while the far corner remains anchored (< 0.3 mm motion). Applying the negated load ([−2.6, −2.6, 0] N)

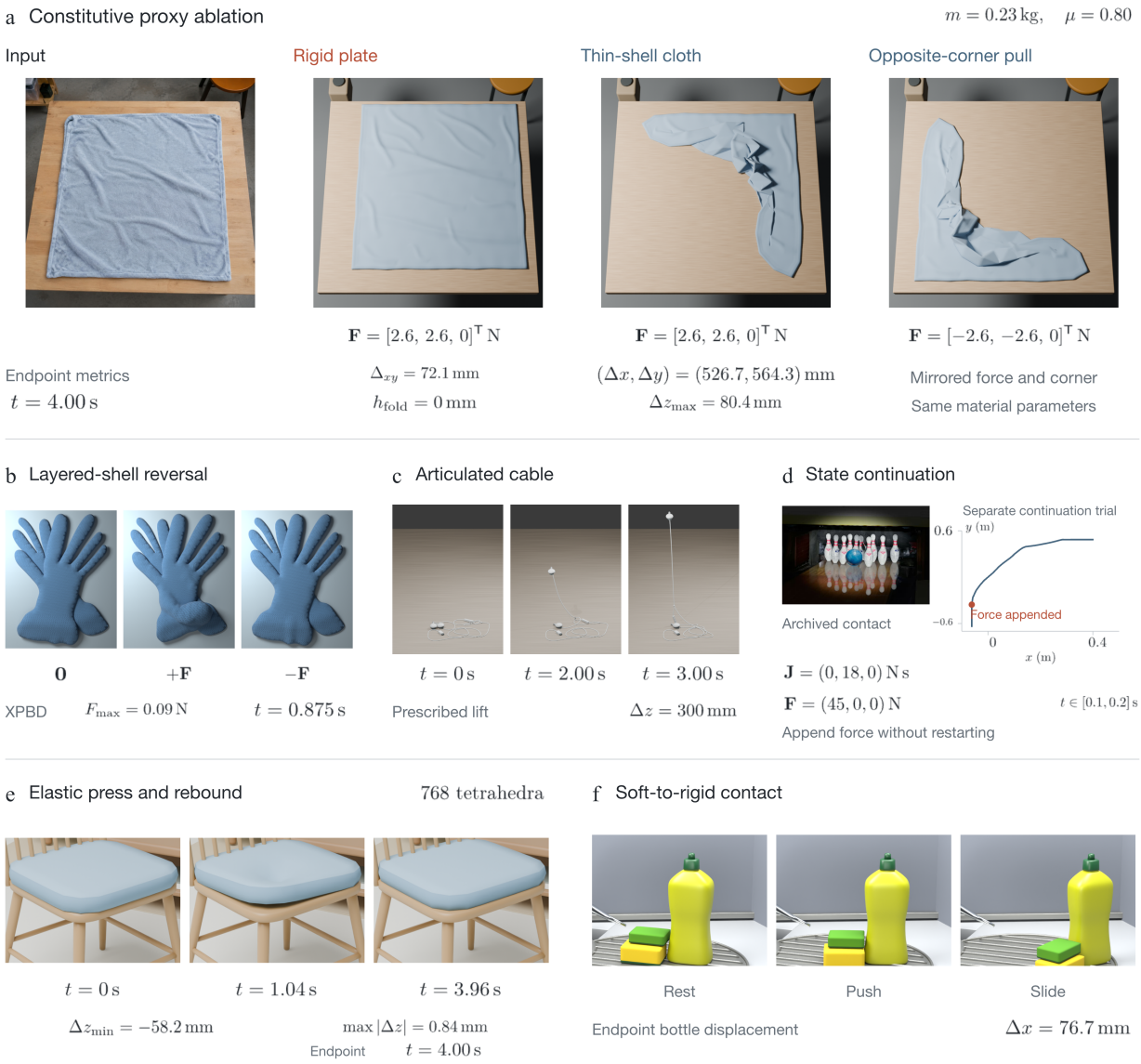


Figure 4: Constitutive proxy ablation, force reversal, and persistent multi-regime state. (a) Controlled ablation on a 0.23 kg microfiber bath towel ($\mu = 0.80$) under $[2.6, 2.6, 0]^\top$ N corner load ($t = 4.00$ s endpoint metrics; frames shown at $t = 3.96$ s): a matched rigid plate slides $\Delta_{xy} = 72.1$ mm in plane with $h_{\mathrm{fold}} = 0$ mm, whereas the 121-node cloth shell pulls $(\Delta x, \Delta y) = (526.7, 564.3)$ mm (771.9 mm planar norm) with $\Delta z_{\max} = 80.4$ mm nodal pleat height (108 mm skinned crest), and mirrored $[-2.6, -2.6, 0]^\top$ N opposite-corner loading produces a symmetric fold. (b) High-resolution 831-vertex XPBD layered-shell glove verification ($F_{\max} = 0.09$ N, shown at $t = 0.875$ s; Appendix A) comparing zero (**0**), forward (**+F**, 76.0 mm peak vertex lift), and reversed (**−F**, 3.87 mm motion) cuff loading, complementing the 2×99-node native finite-element shell in Table 2. (c) Prescribed $\Delta z = 300$ mm lift along a 164-link articulated cable ($t = 0, 2.00, 3.00$ s). (d) Close-up impulse-to-force bowling continuation appending $\mathbf{F} = (45, 0, 0)$ N over $t \in [0.1, 0.2]$ s after an initial $\mathbf{J} = (0, 18, 0)$ N · s impulse without state reset, complementing the full-lane 235.6 N pin-cascade rollout in Table 2. (e) Volumetric indentation ($\Delta z_{\min} = -58.2$ mm at $t = 1.04$ s) and elastic rebound ($\max|\Delta z| = 0.84$ mm at $t = 4.00$ s; frames shown at $t = 3.96$ s) on a 768-tetrahedron seat cushion. (f) Soft-to-rigid contact where two tetrahedral foam sponges compress against a rigid bottle and slide it $\Delta x = 76.7$ mm.

to the diagonally opposite corner patch from the same initial state $\mathbf{s}_0$ produces a mirrored diagonal fold.

Layered Shell Disocclusion and Tension Transmission in an Articulated Cable. Explicit contact and joint constraints likewise isolate actuated branches from passive neighbors in multi-body arrangements (Figure 4b–d). On the stacked knitted wool gloves, upward-oblique cuff loading on the 2×99-node native finite-element

Table 2: Quantitative evaluation across seven scenes and a controlled rigid-plate constitutive ablation. Every configuration is evaluated from a saved equilibrium state $\mathbf{s}_0$ under no-actuation settling (4.0 s horizon for deformable/hybrid scenes and 2.0–3.0 s for rigid/articulated scenes; peak nodal drift is < 1.0 mm across all deformable cases), forward actuation, counterfactual reversal or live continuation, deterministic checkpoint replay ($< 10^{-12}$ state error, 0.000 mm across all cases), and half-timestep ($\Delta t/2$) step-size sensitivity.

Scene & Regime	Discretization	No-Act. Drift	Perturbation Protocol, Counterfactual / Ablation Contrast & $\Delta t/2$ Sensitivity
Ten-pin bowling (*Rigid cascade*)	11 rigid bodies (Δt = 0.17–0.25 ms)	0.00 mm	235.6 N horizontal launch topples 8/10 pins; live mid-rollout continuation alters pin cascade to 9/10 pins (0.614 m terminal ball divergence; 0.53 mm pre-impact $\Delta t/2$ diff).
Dual-pen holder (*Container clearance*)	3 rigid bodies (2.40 mm gap)	$\leq$ 0.15 mm	Axial pulses (0.39 N, 0.59 N) lift target pen 220–662 mm out of cup while unactuated neighbor shifts $< 10^{-15}$ m (0.05 mm airborne $\Delta t/2$ diff).
Wired earphones (*Articulated chain*)	164 capsule links + 5 terminal bodies	$\leq$ 2.83 mm	Prescribed +300 mm left-bud lift raises splitter 121.8 mm and controller 95.0 mm while frictional slack holds right bud to 1.26 mm (1.42 mm RMS $\Delta t/2$ diff).
Bath towel (*Rigid-plate ablation*)	1 rigid plate (0.23 kg, μ = 0.80)	0.11 mm	Identical [2.6, 2.6, 0] N corner tug ($\sin^2$, 0.20–0.36 s) yields 51.2 mm X, 50.8 mm Y flat slab shift (72.1 mm planar norm, 0.0 mm fold; 0.11 mm $\Delta t/2$ RMS).
Bath towel (*Thin-shell cloth*)	121 nodes, 200 tris (0.23 kg, μ = 0.80)	0.15 mm	Same [2.6, 2.6, 0] N tug pulls corner 526.7 mm X, 564.3 mm Y (771.9 mm endpoint planar, 787 mm peak 3D) with 80.4 mm nodal (108 mm skinned) pleats; far corner moves 0.23 mm (20.7 mm $\Delta t/2$ RMS).
Knitted wool gloves (*Layered shells*)	2 × 99 nodes, 320 tris (0.042 kg each)	$\leq$ 0.24 mm	Oblique cuff load ([0.11, 0.025, 0.175] N, 0.25–1.05 s) peels upper glove 386 mm peak while lower moves 0.67 mm; $-\mathbf{F}(t)$ shifts 1.66 mm (1.22 mm $\Delta t/2$ RMS; XPBD shell in Fig. 4b & App. A).
Seat cushion & chair (*3D solid + rigid*)	243 nodes, 768 tets + 4.5 kg chair	$\leq$ 0.41 mm	Downward-oblique press ([3.5, 4.0, −32.0] N, 0.25–1.55 s) indents foam −58.2 mm peak at t = 1.04 s (−50.7 mm at t = 0.90 s) and rebounds to 0.84 mm Z residual at t = 4.00 s (< 0.7 mm $\Delta t/2$ CoM diff).
Sponges & bottle (*3D hybrid contact*)	2 × 36 nodes, 144 tets + 0.33 kg bottle	$\leq$ 0.57 mm	Lateral squeeze ([0.82, 0.04, 0] N + [0.52, 0.03, 0] N, 0.25–0.95 s) compresses sponges against bottle, sliding bottle +76.7 mm upright on wet steel (μ = 0.18; 0.3 mm $\Delta t/2$ X diff).

shell peels the upper glove 38.6 cm across the lower shell while tabletop friction holds the disoccluded lower glove below 0.7 mm drift; reversing the load vector ($-\mathbf{F}(t)$) presses the cuff into unilateral support and limits motion to 1.66 mm (Table 2), a directional contrast corroborated on the 831-vertex compliant XPBD shell in Figure 4b (76.0 mm forward cuff lift vs. 3.87 mm under $-\mathbf{F}$; Appendix A). Along the 164-segment wired earphone assembly, a prescribed +30 cm lift on the left earbud transmits catenary tension to raise the inline controller and Y-splitter by 9.5–12.2 cm, whereas frictional slack beyond the splitter isolates the unactuated right earbud (1.3 mm motion; Figure 4c). In rigid ensembles, the checkpointed solver preserves zero numerical cross-talk ($< 10^{-15}$ m) on an unactuated neighbor across a 2.4 mm pen-holder gap and supports both full-lane pin-cascade continuation (Table 2) and mid-rollout impulse-to-force steering in ten-pin bowling without resetting $\mathbf{s}_0$ (Figure 4d).

Volumetric Elastic Recovery and Soft-to-Rigid Momentum Transfer. Extending from 2D shells to 3D tetrahedral elastodynamics (d = 3) captures through-thickness compression and compliant momentum transfer to movable rigid bodies (Figure 4e–f). Under a downward-oblique load peaking at 32 N vertical force, the 768-tetrahedron seat cushion reaches −5.8 cm peak world-frame nodal indentation at t = 1.04 s atop a dynamic 4.5 kg wooden chair and recovers elastically upon unloading to a 0.84 mm vertical residual at t = 4.00 s. In the hybrid kitchen sink scene, lateral compression across the two tetrahedral foam sponges (2 × 36 nodes, 144 total tetrahedra) transfers horizontal impulse across the compliant contact interface to slide the 0.33 kg detergent bottle +76.7 mm upright along the wet steel drainboard (μ = 0.18), with 0.3 mm horizontal travel sensitivity under $\Delta t/2$ refinement.

Table 3: Quantitative comparison of physics-grounded video synthesis against direct image-to-video baselines. Evaluated across Rigid ($N = 3$), Deformable & Hybrid ($N = 4$), All Primary ($N = 7$), and Counterfactual opposite towel pull ($N = 1$) subsets. Arrows indicate whether higher (↑) or lower (↓) values reflect better performance; $\cos\theta_{\text{dir}}$ measures net motion-centroid directional alignment with the simulated trajectory (excluding the zero-net-translation seat cushion entry on A14B Canny, which averages over 3 deformable and 6 primary scenes).

Subset	Generator & Control	Mask IoU ↑	Traj (%) ↓	PSNR ↑	SSIM ↑	DINOv2 (Full/ROI) ↑	$\cos\theta_{\text{dir}}$ ↑	VRAM	Time (s)
Rigid ($N = 3$)	Direct I2V 1.3B (No Phys)	0.1906	12.58	**25.71**	0.9351	0.8021 / 0.7862	+0.6398	19.4 GB	74.5
	Direct I2V A14B (No Phys)	0.2301	10.40	22.49	0.9520	0.7643 / 0.7210	−0.2053	41.2 GB	354.4
	Wan2.1-VACE 1.3B (Canny)	0.4556	3.84	10.21	0.6918	0.5912 / 0.6174	+0.8161	19.4 GB	55.9
	Wan2.1-VACE 1.3B (RGB)	**0.8424**	**1.00**	16.87	0.8676	0.6502 / 0.5696	+0.9549	19.4 GB	**55.1**
	Wan2.2-VACE A14B (Canny)	0.6936	2.31	22.29	0.9286	0.6416 / 0.6526	+0.3080	41.0 GB	291.4
	Wan2.2-VACE A14B (RGB)	0.6352	3.42	23.46	0.9306	0.6898 / 0.6540	+0.5358	41.0 GB	257.7
	MiniMax-H3 Ref2VA[†]	0.5141	4.05	24.68	**0.9544**	**0.9368 / 0.9134**	−0.0996	35.1 GB	66.4
Deform. ($N = 4$)	Direct I2V 1.3B (No Phys)	0.1617	20.94	**25.59**	0.8596	**0.9607 / 0.9567**	+0.1631	19.4 GB	77.7
	Direct I2V A14B (No Phys)	0.2948	16.65	22.77	**0.9002**	0.8605 / 0.8437	−0.1986	41.2 GB	314.2
	Wan2.1-VACE 1.3B (Canny)	0.4910	4.84	10.80	0.5583	0.8388 / 0.8353	+0.7075	19.4 GB	77.4
	Wan2.1-VACE 1.3B (RGB)	**0.8850**	**0.75**	13.98	0.6689	0.6399 / 0.6260	+0.9212	19.4 GB	**77.2**
	Wan2.2-VACE A14B (Canny)	0.5659	3.80	18.41	0.7046	0.8365 / 0.8264	+0.9293	41.0 GB	296.8
	Wan2.2-VACE A14B (RGB)	0.6002	2.69	17.89	0.6945	0.8202 / 0.8041	+0.6540	41.0 GB	292.2
Primary ($N = 7$)	Direct I2V 1.3B (No Phys)	0.1741	17.35	**25.64**	0.8919	**0.8927 / 0.8836**	+0.3674	19.4 GB	76.3
	Direct I2V A14B (No Phys)	0.2670	13.97	22.65	**0.9224**	0.8193 / 0.7911	−0.2015	41.2 GB	331.4
	Wan2.1-VACE 1.3B (Canny)	0.4758	4.41	10.55	0.6155	0.7327 / 0.7419	+0.7540	19.4 GB	68.2
	Wan2.1-VACE 1.3B (RGB)	**0.8667**	**0.86**	15.22	0.7541	0.6443 / 0.6019	+0.9356	19.4 GB	**67.7**
	Wan2.2-VACE A14B (Canny)	0.6206	3.16	20.07	0.8006	0.7530 / 0.7519	+0.6187	41.0 GB	294.5
	Wan2.2-VACE A14B (RGB)	0.6152	3.00	20.28	0.7957	0.7643 / 0.7398	+0.6033	41.0 GB	277.4
Opposite (*Towel*)	Direct I2V 1.3B (No Phys)	0.0043	34.30	**30.15**	**0.9164**	**0.9900 / 0.9878**	−0.7441	19.4 GB	69.7
	Direct I2V A14B (No Phys)	0.4444	18.51	25.49	0.9012	0.7890 / 0.7898	−0.1702	41.2 GB	347.9
	Wan2.1-VACE 1.3B (Canny)	0.3452	16.56	9.49	0.4369	0.7028 / 0.6956	+0.9930	19.4 GB	**69.0**
	Wan2.1-VACE 1.3B (RGB)	**0.9697**	**0.19**	12.00	0.5828	0.3302 / 0.3268	+0.9944	19.4 GB	69.4
	Wan2.2-VACE A14B (Canny)	0.5713	13.30	17.15	0.6813	0.5130 / 0.5214	+0.9579	41.0 GB	356.4
	Wan2.2-VACE A14B (RGB)	0.5680	10.34	16.75	0.6355	0.5265 / 0.5305	+0.9083	41.0 GB	337.7

[†]Evaluated on the three rigid benchmark scenes at 768 × 768 (124 frames, 24 fps, 8×H100 FSDP/Ulysses, peak VRAM/GPU) with 32 kHz stereo audio.

4.3 Physics-Grounded Video Synthesis vs. Direct Image-to-Video Generation

We evaluate generative video grounding across seven primary interactions—ceramic mug tumbling, kraft notebook sliding, vegetable-bowl collision, bath towel corner pulling, knitted wool glove peeling, sponge-to-bottle pushing, and seat cushion indentation—together with two counterfactual queries testing opposite-corner towel pulling across all six VACE and Direct I2V variants and horizontal mug sliding across the two direct baselines (53 generated videos in total; Table 3 and Appendix B). Four trajectory-conditioned VACE variants spanning Wan2.1-VACE 1.3B and Wan2.2-VACE-Fun A14B under Φ_{Canny} and Φ_{RGB} (512×512, 81 frames, 30 denoising steps) are compared against image-and-text-conditioned Direct I2V 1.3B and A14B baselines anchored on the same input photograph I_0 and action prompts without V_{phys}, alongside rigid-scene MiniMax-H3 Ref2VA rollouts. Motion metrics compare each generated video against temporally resampled physics rollouts via framewise dynamic motion-mask IoU, normalized motion-centroid trajectory error (as image-width percentage), and net start-to-end motion-direction cosine alignment $\cos\theta_{\text{dir}}$, while appearance metrics compare generated frames against I_0 via static-background PSNR and SSIM and full-frame/ROI DINOv2 ViT-L/14-reg4 [7, 33] cosine similarity.

Failure of Direct I2V Without Trajectory Conditioning on Stateful and Counterfactual Actions. Table 3 and Figure 5 expose two contrasting failure patterns in direct image-to-video baselines lacking 3D trajectory grounding: static frame copying at 1.3B and ungrounded motion hallucination at A14B. Direct I2V 1.3B attains high static PSNR (> 25 dB) and DINOv2 similarity (0.89) primarily by leaving deformable scenes nearly frozen: its dynamic motion-mask IoU falls below 0.03 on both the bath towel and the seat cushion, yielding > 20% mean centroid trajectory error across deformable and hybrid scenes. Scaling to Direct I2V A14B increases pixel motion but still lacks 3D contact or force grounding: it levitates the ceramic mug mid-air, expands the red cabbage in place rather than rolling it into the bowl, and morphs glove fingers without peeling the upper shell. Under counterfactual cloth prompts, Direct I2V A14B folds the lower-right towel corner regardless

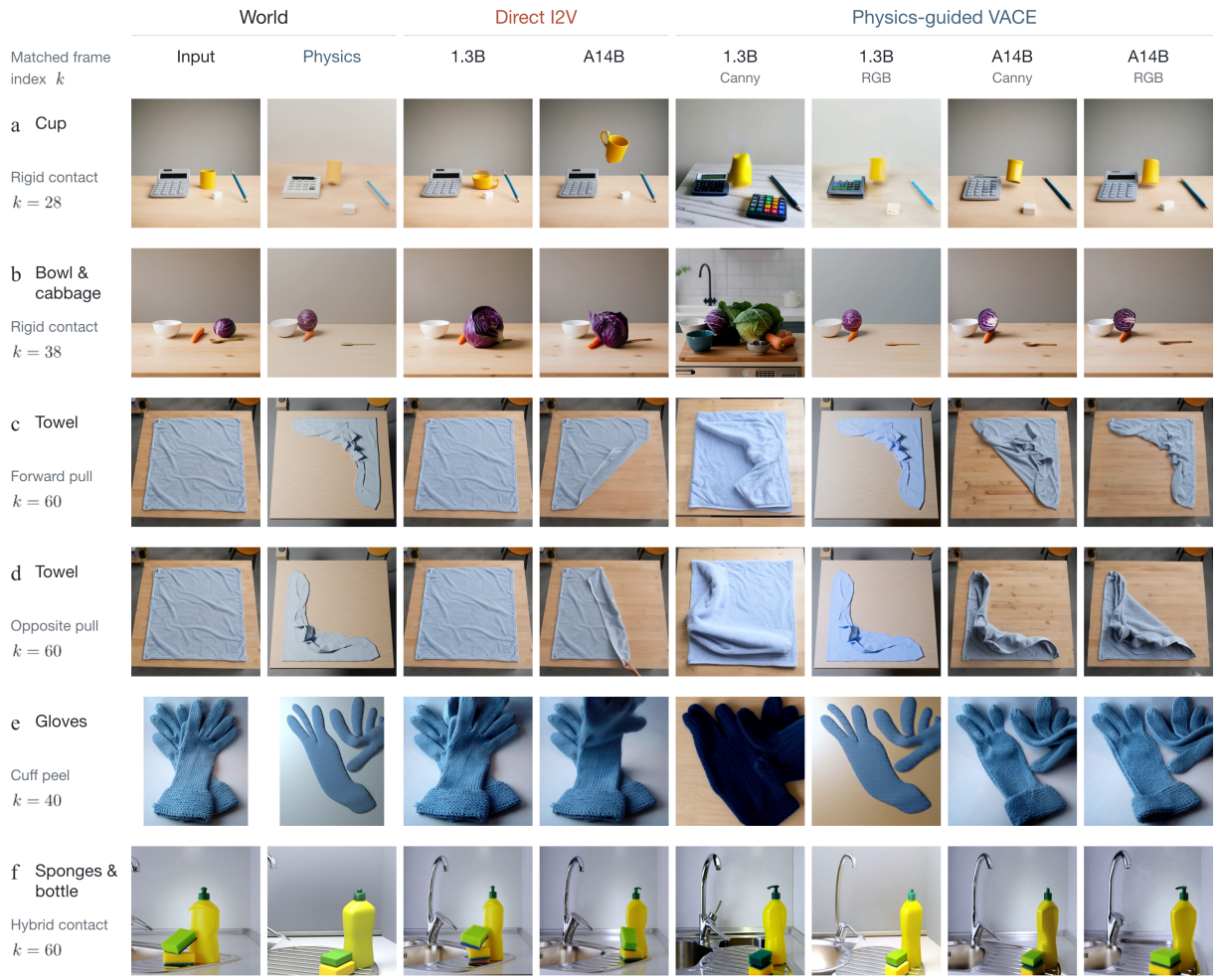


Figure 5: Phase-aligned comparison of physics-grounded video synthesis against direct image-to-video baselines. Rows (a–f) span ceramic cup tumbling, bowl-and-cabbage collision, forward and counterfactual opposite-corner towel pulling, knitted glove cuff peeling, and hybrid sponge-to-bottle pushing. Columns compare the input photograph and aligned physics rollout against image-and-text-only Direct I2V (1.3B, A14B) and trajectory-conditioned VACE (1.3B and A14B under Canny and RGB control) at matched generated frame indices $(28, 38, 60, 60, 40, 60)$ for rows (a–f), corresponding to temporally resampled physics frames $(42, 57, 71, 71, 48, 71)$.

of whether the prompt specifies pulling the lower-left or upper-right corner, producing negative directional alignment ($\cos\theta_{\text{dir}} \approx -0.20$ across primary scenes and -0.17 on the opposite pull). Across the seven primary scenes, conditioning VACE on simulated 3D rollouts raises mean motion-mask IoU by 2.3×–5.0× and cuts mean centroid trajectory error by 3.9×–20.2× relative to the direct baseline of the same model scale, while restoring $>$ +0.90 directional cosine alignment across both forward and counterfactual cloth pulls.

Backbone Capacity and Conditioning Modality Trade-Off. Comparing Φ_{Canny} and Φ_{RGB} across the 1.3B backbone and the A14B Mixture-of-Experts backbone (14B active parameters) clarifies how structural control density interacts with denoiser capacity. On the 1.3B backbone, sparse Canny contours (Φ_{Canny}) leave background illumination and interior appearance weakly constrained, lowering mean static-background fidelity to 10.6 dB PSNR with visible tabletop texture drift (Figure 5, column 5). Dense RGB renders (Φ_{RGB}) on 1.3B achieve high silhouette alignment with the simulated trajectory (mean IoU 0.87, trajectory error $<$ 1%), but cause the compact denoiser to copy flat synthetic mesh shading onto deformable textiles, reducing mean DINOv2 similarity to 0.64 on deformable scenes. Scaling to Wan2.2-VACE-Fun A14B raises mean primary-set

static-background PSNR by +9.5 dB under Canny control and +5.1 dB under RGB control relative to the corresponding 1.3B configuration, while lifting mean deformable and hybrid DINOv2 similarity to 0.82–0.84 (full-frame) and 0.80–0.83 (ROI) at 2.1× higher peak GPU memory (41.0 GB vs. 19.4 GB). On rigid scenes, dual-reference MiniMax-H3 Ref2VA complements pixel-aligned VACE control by preserving 768^2 photographic reflectance (0.94 DINOv2 similarity, 24.7 dB static PSNR) and synthesizing synchronized 32 kHz impact audio with looser pixel-level trajectory locking.

5 Conclusion

We presented TelePhysica, a closed-loop framework that reconstructs executable, multi-regime 3D physical worlds from a single photograph to drive stateful interaction and trajectory-grounded video synthesis. Pairing pixel-aligned visual critique with a zero-force gravitational equilibrium preflight raises verified single-image scene recovery from 0% in open-loop synthesis to 81% across 100 tabletop scenes, while unified rigid, articulated, thin-shell, and 3D tetrahedral elastodynamics captures localized cloth folding, elastic foam rebound, and soft-to-rigid momentum transfer for controllable video diffusion. Constitutive parameters such as mass density, Young's modulus, and friction coefficients are currently assigned from material priors rather than identified online from multi-step visual interaction feedback; estimating these parameters from observed interaction responses remains a natural direction for future work.

A Mathematical Formulation, Simulation Conventions, and Verification Criteria

Coordinate Frame, Units, and Pose Conventions. All scene scripts, solvers, and perturbation schedules share a right-handed Cartesian world frame with $+X$ right, $+Y$ forward in depth, and $+Z$ vertically upward against gravity $\mathbf{g} = (0, 0, -9.81)^\top \,\mathrm{m/s^2}$. Lengths are expressed in meters (m), masses in kilograms (kg), time in seconds (s), forces in newtons (N), moments in newton-meters (N · m), linear impulses in newton-seconds (N · s), and elastic moduli in pascals (Pa). Rigid-body orientations use scalar-first unit quaternions $\mathbf{q}_i^{\mathrm{rot}} = (q_w, q_x, q_y, q_z) \in S^3$ ($\|\mathbf{q}_i^{\mathrm{rot}}\|_2 = 1$) mapped to orthogonal matrices $\mathbf{R}(\mathbf{q}_i^{\mathrm{rot}}) \in \mathrm{SO}(3)$ when composing spatial transforms $\mathbf{T} \in \mathrm{SE}(3)$.

Native Finite-Element Flex Compilation and Compliant XPBD Thin Shells. In our declarative schema, deformable objects specify persistent node identifiers, world-space rest coordinates $\bar{\mathbf{X}} \in \mathbb{R}^{n_v \times 3}$, optional prestressed initial coordinates $\mathbf{X}(0) \in \mathbb{R}^{n_v \times 3}$, homogeneous element connectivity $\mathcal{E}$ using triangles for $d = 2$ shells and tetrahedra for $d = 3$ solids, pinned anchor subsets, and constitutive parameters $(m, E, \nu, \eta, \mu, r_c)$. During native MuJoCo flexible-body compilation [38], world-space nodes map into the parent body frame via $\mathbf{x}_a^{\mathrm{loc}} = \mathbf{R}_0^\top (\mathbf{x}_a - \mathbf{p}_0)$, and lumped nodal masses distribute across vertex bodies carrying three translational degrees of freedom per free node. Implicit-in-velocity integration then advances membrane/bending ($d = 2$) or volumetric continuum ($d = 3$) elastic forces jointly with unilateral frictional contact constraints.

For an 831-vertex high-resolution knitted wool glove shell (499 front-glove and 332 rear-glove midsurface vertices, 0.035 kg per glove, $\Delta t = 1/480$ s), we additionally verify Extended Position-Based Dynamics (XPBD) edge-stretch and cross-edge distance constraints [29, 31] with edge-stretch stiffness $k_{\mathrm{stretch}} = 120.0\,\mathrm{N/m}$, cross-edge bending stiffness $k_{\mathrm{bend}} = 0.8\,\mathrm{N/m}$, compliance $\alpha_c = 1/k$, and timestep-scaled compliance $\tilde{\alpha} = 1/(k\Delta t^2)$ (units kg^{-1}, with $m_a^{-1} = 0$ for pinned nodes). Within each Gauss-Seidel substep iteration, Lagrange multiplier increments update nodal pairs (a, b) with rest length L_{ab}^0 along $\hat{\mathbf{n}}_{ab} = (\tilde{\mathbf{x}}_a - \tilde{\mathbf{x}}_b)/\|\tilde{\mathbf{x}}_a - \tilde{\mathbf{x}}_b\|_2$ according to:

$$\Delta\lambda_{ab} = -\frac{\|\tilde{\mathbf{x}}_a - \tilde{\mathbf{x}}_b\|_2 - L_{ab}^0 + \tilde{\alpha}\,\lambda_{ab}}{m_a^{-1} + m_b^{-1} + \tilde{\alpha}}, \qquad \Delta\tilde{\mathbf{x}}_a = +m_a^{-1}\Delta\lambda_{ab}\,\hat{\mathbf{n}}_{ab}, \qquad \Delta\tilde{\mathbf{x}}_b = -m_b^{-1}\Delta\lambda_{ab}\,\hat{\mathbf{n}}_{ab}. \tag{7}$$

Over a 3.0 s zero-input settle, center-of-mass drift is 0.012 mm (0.226 mm peak vertex drift). A localized 0.09 N front-cuff $\sin^2$ load pulse ($t \in [0.35, 0.95]$ s) produces 75.99 mm peak vertex lift (10.59 mm CoM shift), reversed loading $-\mathbf{F}(t)$ yields 3.87 mm peak vertex motion against tabletop friction ($\mu = 0.55$), deterministic replay error is 0.000 mm, and half-step refinement ($\Delta t/2 = 1/960$ s) gives a 3.32 mm terminal vertex RMS difference (4.37% of peak).

B Per-Scene Video Synthesis Quantitative Ledger

Table 5 reports the complete per-scene quantitative evaluation across all seven primary multi-regime scenes and two counterfactual action queries (53 evaluated video generations across Direct I2V 1.3B, Direct I2V A14B, Wan2.1-VACE 1.3B Canny/RGB, Wan2.2-VACE-Fun A14B Canny/RGB, and MiniMax-H3 Base Ref2VA). On the upholstered seat cushion scene, downward indentation is followed by elastic recovery to within 0.84 mm of the initial surface height at $t = 4.0$ s; start-to-end 2D centroid translation is therefore near zero by physical design, whereas framewise dynamic motion-mask IoU and trajectory error capture the transient press-and-rebound cycle.

Table 4: Automated geometric and physical verification gates enforced across scene reconstruction and rollout evaluation.

Verification Stage	Evaluated Quantity	Quantitative Acceptance Criterion
Role Inventory	Mobility role ρ_i	Every scene object declares an explicit role; movable foreground bodies and deformables are assigned `dynamic`.
Mesh Topology	Edge incidence & area	Zero non-manifold edges ($\leq$ 2 faces per edge) and zero degenerate triangles below area $10^{-10}\ \max(L_{\text{bbox}}, 1\,\text{m})^2$, where L_{bbox} is the maximum bounding-box extent.
Stage-1 Visual Gate	Silhouette & edge IoU	Combined occupancy, 72-bin radial contour profile, and Sobel interior edge score ≥ 0.45 with foreground mask IoU ≥ 0.45 and silhouette structure ≥ 0.60.
Stage-2 Equilibrium	Initial contact & 1.0 s settle	Pairwise initial overlap $\delta_{ij}(0) \leq \varepsilon_{c,ij} = \max(0.1\,\text{mm}, 0.01 \min_{k\in\mathcal{D}_{ij}} \ell_k)$ (ℓ_k: smallest declared collider thickness or contact diameter of dynamic object k in pair), drift $\leq \varepsilon_{\text{eq},i} = \max(1.0\,\text{mm}, 0.02\,\ell_i)$, terminal speed $\leq \varepsilon_{v,i}$ ($\max(0.01\,\text{m/s}, 0.05\,\text{s}^{-1}\ell_i)$ rigid; $\max(0.01\,\text{m/s}, 0.05\,\text{s}^{-1}\varepsilon_{\text{eq},i})$ deformable), rigid rotation $\leq 5°$, angular speed ≤ 0.1 rad/s, and 0 solver warnings.
Perturbation Rollout	Multi-trial execution	Schema-validated interventions execute across forward, negated counterfactual ($-\mathbf{F}$), replay, and $\Delta t/2$ trials with finite integration state, 0 solver warnings, and logged force-time integrals $\int \mathbf{F}(t)\,\mathrm{d}t$ and $\Delta t/2$ RMS sensitivity.
State Determinism	Checkpoint & continuation	Full-state checkpoint replay, midpoint pause/restore continuation, and chunked stepping reproduce complete integration state vectors within $\leq 10^{-12}$ maximum componentwise error under identical hardware/software execution.

Table 5: Complete per-scene quantitative video generation ledger across rigid, thin-shell deformable, 3D hybrid, and counterfactual evaluations.

Scene (Regime)	Model & Conditioning	Mask IoU ↑	Traj Err (%) ↓	PSNR (dB) ↑	SSIM ↑	DINOv2 (Full / ROI) ↑	$\cos\theta_{\text{dir}}$ ↑	Time (s)
Yellow Ceramic Mug (*Rigid tumble*)	Direct I2V 1.3B	0.1004	13.57	30.49	0.9660	0.9525 / 0.9292	+0.3515	69.0
	Direct I2V A14B	0.2070	8.99	24.26	0.9613	0.8562 / 0.8319	−0.5519	376.1
	VACE 1.3B (Canny)	0.4968	3.63	11.14	0.7200	0.8253 / 0.7989	+0.9079	55.8
	VACE 1.3B (RGB)	0.7125	2.07	12.04	0.8397	0.7480 / 0.6426	+0.9953	56.0
	VACE A14B (Canny)	0.6194	3.64	18.46	0.8747	0.8738 / 0.7668	+0.5897	253.6
	VACE A14B (RGB)	0.4334	7.64	18.90	0.8746	0.8367 / 0.7964	+0.9296	245.4
	MiniMax-H3 Ref2VA	0.1888	9.28	24.48	0.9662	0.9542 / 0.9373	−0.7232	66.4
Kraft Notebook & Pencils (*Rigid slide*)	Direct I2V 1.3B	0.2268	8.59	23.11	0.9395	0.7480 / 0.7520	+0.6178	74.5
	Direct I2V A14B	0.2322	4.98	21.90	0.9477	0.6695 / 0.5921	−0.8377	314.2
	VACE 1.3B (Canny)	0.4919	1.18	8.43	0.6735	0.2945 / 0.3513	+0.5428	55.6
	VACE 1.3B (RGB)	0.8779	0.68	18.70	0.8222	0.5024 / 0.4667	+0.8693	55.0
	VACE A14B (Canny)	0.7091	1.98	21.04	0.9329	0.5220 / 0.6715	−0.6634	301.1
	VACE A14B (RGB)	0.7384	1.24	22.81	0.9427	0.7470 / 0.6826	−0.3003	251.7
	MiniMax-H3 Ref2VA	0.6060	2.03	24.47	0.9350	0.9443 / 0.8798	−0.5745	66.3
Bowl & Red Cabbage (*Rigid collision*)	Direct I2V 1.3B	0.2446	15.58	23.52	0.8997	0.7057 / 0.6774	+0.9500	79.9
	Direct I2V A14B	0.2510	17.23	21.31	0.9471	0.7673 / 0.7389	+0.7736	372.9
	VACE 1.3B (Canny)	0.3782	6.71	11.05	0.6820	0.6539 / 0.7021	+0.9977	56.3
	VACE 1.3B (RGB)	0.9369	0.25	19.88	0.9409	0.7003 / 0.5996	+1.0000	54.4
	VACE A14B (Canny)	0.7524	1.32	27.37	0.9783	0.5290 / 0.5194	+0.9976	319.5
	VACE A14B (RGB)	0.7337	1.37	28.68	0.9746	0.4858 / 0.4830	+0.9782	276.1
	MiniMax-H3 Ref2VA	0.7476	0.86	25.09	0.9619	0.9118 / 0.9232	+0.9989	66.4
Microfiber Bath Towel (*Thin-shell forward*)	Direct I2V 1.3B	0.0132	26.16	30.70	0.9263	0.9893 / 0.9877	+0.0661	76.0
	Direct I2V A14B	0.2458	26.97	27.50	0.9265	0.7267 / 0.7304	+0.2334	308.6
	VACE 1.3B (Canny)	0.3897	5.95	9.73	0.4872	0.6315 / 0.6425	+0.9503	79.8
	VACE 1.3B (RGB)	0.9671	0.37	12.50	0.5931	0.3390 / 0.3395	+0.9973	75.0
	VACE A14B (Canny)	0.7278	1.09	20.36	0.7059	0.6050 / 0.6135	+0.9300	275.7
	VACE A14B (RGB)	0.7957	2.37	18.51	0.6545	0.5245 / 0.5078	+0.9073	278.0
Stacked Knitted Gloves (*Layered shell peel*)	Direct I2V 1.3B	0.3479	6.07	18.38	0.7064	0.8757 / 0.8743	+1.0000	80.7
	Direct I2V A14B	0.3274	15.32	18.30	0.8094	0.8648 / 0.8653	−0.7941	328.6
	VACE 1.3B (Canny)	0.3799	9.89	9.48	0.5710	0.8297 / 0.8279	+0.7946	78.5
	VACE 1.3B (RGB)	0.9338	0.44	14.14	0.7103	0.4412 / 0.4525	+0.9904	81.2
	VACE A14B (Canny)	0.5749	5.24	21.34	0.7591	0.8814 / 0.8877	+0.8581	310.7
	VACE A14B (RGB)	0.5921	2.28	20.97	0.7540	0.8851 / 0.8913	+0.9791	306.4
Kitchen Sponges & Bottle (*3D hybrid contact*)	Direct I2V 1.3B	0.2615	17.10	24.21	0.8538	0.9816 / 0.9736	+0.5269	70.5
	Direct I2V A14B	0.5158	8.05	24.66	0.9499	0.9199 / 0.9286	−0.6672	299.2
	VACE 1.3B (Canny)	0.8181	1.60	14.93	0.7251	0.9367 / 0.9331	+0.9998	70.8
	VACE 1.3B (RGB)	0.8720	1.44	14.83	0.7513	0.8235 / 0.7929	+0.7557	69.6
	VACE A14B (Canny)	0.7361	3.19	19.10	0.8058	0.9421 / 0.9340	+0.9999	283.6
	VACE A14B (RGB)	0.7903	1.23	18.51	0.7969	0.9461 / 0.9331	+0.9947	279.7
Seat Cushion & Chair (*3D elastic rebound*)	Direct I2V 1.3B	0.0241	34.42	29.07	0.9518	0.9961 / 0.9913	−0.9408	83.4
	Direct I2V A14B	0.0900	16.26	20.60	0.9150	0.9304 / 0.8505	+0.4334	320.3
	VACE 1.3B (Canny)	0.3762	1.91	9.06	0.4500	0.9573 / 0.9377	+0.0852	80.5
	VACE 1.3B (RGB)	0.7669	0.75	14.43	0.6211	0.9557 / 0.9193	+0.9415	83.1
	VACE A14B (Canny)	0.2247	5.67	12.85	0.5474	0.9174 / 0.8705	—	317.1
	VACE A14B (RGB)	0.2228	4.89	13.55	0.5724	0.9252 / 0.8842	−0.2653	304.6
Microfiber Bath Towel (*Counterfactual opposite*)	Direct I2V 1.3B	0.0043	34.30	30.15	0.9164	0.9900 / 0.9878	−0.7441	69.7
	Direct I2V A14B	0.4444	18.51	25.49	0.9012	0.7890 / 0.7898	−0.1702	347.9
	VACE 1.3B (Canny)	0.3452	16.56	9.49	0.4369	0.7028 / 0.6956	+0.9930	69.0
	VACE 1.3B (RGB)	0.9697	0.19	12.00	0.5828	0.3302 / 0.3268	+0.9944	69.4
	VACE A14B (Canny)	0.5713	13.30	17.15	0.6813	0.5130 / 0.5214	+0.9579	356.4
	VACE A14B (RGB)	0.5680	10.34	16.75	0.6355	0.5265 / 0.5305	+0.9083	337.7
Ceramic Mug (*Counterfactual*)	Direct I2V 1.3B	0.0082	20.84	32.12	0.9690	0.9890 / 0.9660	+0.1234	79.4
	Direct I2V A14B	0.2233	15.80	22.03	0.9590	0.8883 / 0.8301	+0.3867	316.5

References

[1] Alibaba PAI. VideoX-Fun: High-performance video generation pipeline for Wan-VACE and dual-expert models. https://github.com/aigc-apps/VideoX-Fun, 2025.

[2] Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, Chi Zhang, Hongyuan Zhang, Wenhao Zhuang, and Xuelong Li. AI Flow: Perspectives, Scenarios, and Approaches. *Vicinagearth*, 3:1, 2026.

[3] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4D-fy: Text-to-4D generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7996–8006, 2024.

[4] 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 (CVPR)*, pages 6178–6189, 2025.

[5] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. DreamScene4D: Dynamic multi-object scene generation from monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[6] Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016.

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[8] Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 18225–18250, 2023.

[9] Yutao Feng, Yintong Shang, Xuan Li, Tianjia Shao, Chenfanfu Jiang, and Yin Yang. PIE-NeRF: Physics-based interactive elastodynamics with NeRF. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4450–4461, 2024.

[10] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.19386.

[11] Guangzhao He, Rundong Luo, Wei-Chiu Ma, and Hadar Averbuch-Elor. Thinking in Blender: Staged executable inverse graphics with vision-language models. *arXiv preprint arXiv:2606.02580*, 2026.

[12] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A. Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An LLM agent for synthesizing 3D scenes as Blender code. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pages 19252–19282, 2024.

[13] Hong Huang, Yuqiu Liu, Chenyu You, Daniel F. Martin, Chuhang Zou, and Wuyang Chen. Fluid-Gen-Zero: Grounding pretrained video generators in physics without training. *arXiv preprint arXiv:2610.10984*, 2026.

[14] Ian Huang, Guandao Yang, and Leonidas J. Guibas. BlenderAlchemy: Editing 3D graphics with vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 297–314, 2024.

[15] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17191–17202, 2025.

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[17] Rui Li, Ke Hao, Yuanzhi Liang, Haibin Huang, Chi Zhang, Yun Gu, and Xuelong Li. Learning to Credit the Right Steps: Objective-aware Process Optimization for Visual Generation. *arXiv preprint arXiv:2604.19234*, 2026.

[18] Rui Li, Bingyu Li, Yuanzhi Liang, Haibin Huang, Chi Zhang, and Xuelong Li. Reward-Aware Trajectory Shaping for Few-step Visual Generation. *arXiv preprint arXiv:2604.14910*, 2026.

[19] Rui Li, Yuanzhi Liang, Ziqi Ni, Haibin Huang, Chi Zhang, and Xuelong Li. Rethinking Reward Signals in Video GRPO: When Scores Become Targets. *arXiv preprint arXiv:2511.19356*, 2026.

[20] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[21] Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. WonderPlay: Dynamic 3D scene generation from a single image and actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9080–9090, 2025.

[22] Jiarong Liang, Max Ku, Ka-Hei Hui, Ping Nie, and Wenhu Chen. VisPhyWorld: Probing physical reasoning via code-driven video reconstruction. *arXiv preprint arXiv:2602.13294*, 2026.

[23] Yuanzhi Liang, Yijie Fang, Rui Li, Ziqi Ni, Ruijie Su, and Chi Zhang. Integrating Reinforcement Learning with Visual Generative Models: Foundations and Advances. *Vicinagearth*, 3:2, 2026.

[24] Yuanzhi Liang, Xuan'er Wu, Yirui Liu, Yijie Fang, Yizhen Fan, Ke Hao, Rui Li, Ruiying Liu, Ziqi Ni, Peng Yu, Yanbo Wang, Haibin Huang, Qizhen Weng, Chi Zhang, and Xuelong Li. TeleBoost: A Systematic Alignment Framework for High-Fidelity, Controllable, and Robust Video Generation. *arXiv preprint arXiv:2602.07595*, 2026.

[25] Jiajing Lin, Zhenzhong Wang, Dejun Xu, Shu Jiang, Yunpeng Gong, and Min Jiang. Phys4DGen: Physics-compliant 4D generation with multi-material composition perception. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

[26] Ruiying Liu, Yuanzhi Liang, Haibin Huang, Tianshu Yu, and Chi Zhang. Learning What to Trust: Bayesian Prior-Guided Optimization for Visual Generation. *arXiv preprint arXiv:2511.18919*, 2025.

[27] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 360–378, 2024.

[28] Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. RealWonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

[29] Miles Macklin, Matthias Müller, and Nuttapong Chentanez. XPBD: Position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games (MIG)*, pages 49–54, 2016.

[30] MiniMax AI. MiniMax-H3: Multimodal reference-to-video (Ref2VA) generation. https://github.com/MiniMax-AI/MiniMax-H3, 2026.

[31] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2):109–118, 2007.

[32] Ziqi Ni, Yuanzhi Liang, Rui Li, Yi Zhou, Haibin Huang, Chi Zhang, and Xuelong Li. Seeing What Matters: Visual Preference Policy Optimization for Visual Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27260–27269, 2026.

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[34] Jiawei Shao and Xuelong Li. AI Flow at the Network Edge. *IEEE Network*, 40(1):330–336, 2026.

[35] Ruihong Shen, Žiga Kovačič, Peter Kulits, Xingrui Wang, Zizhang Li, Joshua B. Tenenbaum, Alan Yuille, Jieneng Chen, and Jiajun Wu. 4DCodeBench: Benchmarking agents on inverse graphics of dynamic scenes. *arXiv preprint arXiv:2610.03715*, 2026.

[36] Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3D-GPT: Procedural 3D modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023.

[37] Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. PhysMotion: Physics-grounded dynamics from a single image. *arXiv preprint arXiv:2411.17189*, 2024.

[38] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

[39] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[40] Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2509.20358.

[41] Hanyang Wang, Yimo Cai, Weiliang Chen, Jiawei Chi, Haowen Sun, Qiyu Dai, Yi-Hsin Hung, Xingzhuo Guo, Jinshan Ren, Runmao Yao, et al. Code as worlds: Agentic discovery of executable world representations for physical reasoning. *arXiv preprint arXiv:2608.27549*, 2026.

[42] 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.

[43] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion control for anything using entity representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–348, 2024.

[44] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4389–4398, 2024.

[45] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3D embodied AI environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16227–16237, 2024.

[46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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

[48] Cheng Yuan, Zhenyu Jia, Jiawei Shao, and Xuelong Li. Enhancing Neural Video Compression of Static Scenes with Positive-Incentive Noise. *arXiv preprint arXiv:2603.06095*, 2026.

[49] Jiahao Zhan, Zizhang Li, Hong-Xing Yu, and Jiajun Wu. PerpetualWonder: Long-horizon action-conditioned 4D scene generation. *arXiv preprint arXiv:2602.04876*, 2026.

[50] Chi Zhang, Yuanzhi Liang, Xi Qiu, Fangqiu Yi, and Xuelong Li. VAST 1.0: A Unified Framework for Controllable and Consistent Video Generation. *arXiv preprint arXiv:2412.16677*, 2024.

[51] Tianyuan Zhang, Hong-Xing Yu, Rundi 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 388–406, 2024.

[52] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2063–2073, 2025.

[53] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3D gaussians. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 407–423, 2024.