Show the Futures the Evidence Supports

Set-valued physical counterfactual editing with Agentic Vision

Abstract

A source clip can be compatible with multiple physical scene models that imply different outcomes after the same edit. We propose a multi-round Agentic Vision system that identifies these ambiguities, tests candidate scene programs against the source, and generates distinct counterfactual branches when their differences are consequential. We will evaluate whether branch-aware editing improves outcome coverage at fixed generation cost while preserving the source prefix and avoiding spurious branches on identifiable cases.

Keywords / agentic vision, counterfactual video editing, identifiability, Blender, rigid-body simulation

Problem and prior work

Two blocks may remain still on a shallow ramp in a source clip, revealing little about their friction coefficients. A request to steepen the ramp can produce either sticking or sliding. VideoPhysEdit [1] reconstructs executable rigid-body scenes and uses simulated trajectories to guide a counterfactual video edit. What-If World [2] tests sensitivity to paired interventions. This proposal targets the cases where the source cannot identify a unique response to the requested intervention. Its unit of evaluation is the set of supported outcome classes and how that set contracts with additional evidence. VideoCoCo [3] provides another precedent for using executable Blender drafts as generation controls.

Unified understanding and generation loop

A multimodal controller reads source video and edit instructions. It uses frame seeking, segmentation, tracking and optical flow to extract constraints, then writes candidate Blender scene programs that share geometry and differ only in selected latent parameters. Python filters candidates by their reprojection error and observed contact events. Simulate the requested edit on the retained programs and group them by qualitative outcome, such as sticking, sliding or toppling. The controller inspects previews and, when supplied an optional diagnostic clip, revises the hypothesis set using the same evidence interface.

Controlled rendering

Render one representative per outcome class using shared camera and assets, and condition a frozen editor on branch-specific trajectories and masks. Keep pre-intervention footage fixed by compositing, then verify branch outcomes and check for unintended changes to unaffected regions. A failed visual check triggers local regeneration within a capped budget. Output a single clip when all surviving programs agree on the outcome, and label alternatives with their physical assumptions when they diverge.

Validation plan

Experimental design

Create 120 procedurally generated scenarios with known parameter ranges, paired factual and edited rollouts, and a diagnostic source clip where available. Use friction thresholds, mass-dependent collisions and support stability. Compute a reference outcome set by dense sweeps with adaptive sampling near transitions, and test sensitivity to grid resolution. Hold out object shapes and intervention magnitudes. Compare point-estimate simulation editing, same-budget video resampling, and unfiltered parameter branching. Include a known-parameter oracle and a shared-renderer comparison.

Outcomes and rejection criteria

Primary outcomes are recall and precision of feasible outcome classes in final videos at matched render budget. Measure excess branching on identifiable cases, source-prefix fidelity, and whether extra evidence removes incompatible branches. Independent annotators score final-video events with no access to method labels. Use scene-clustered intervals. The hypothesis fails if additional branches merely repeat one visible outcome, if they primarily increase false outcomes, or if evidence fails to resolve ambiguity in designed identifiable cases.

Feasibility and deliverables

Weeks 1-2 build synthetic scenario families and parameter sweeps. Weeks 3-4 implement proposal filtering and branch rendering. Weeks 5-7 evaluate generated clips and release programs, ground-truth outcome sets and branch traces. Restrict the first study to a small number of physical unknowns and known camera geometry. If real source clips lack calibration, keep them as qualitative case studies outside the primary test.

References

[1] VideoPhysEdit: Physical Counterfactual Video Editing via Rigid-Body Physical Scene Reconstruction. arXiv:2609.35134.

[2] What-If World: A Causal Benchmark for General World Models in Embodied Scenarios. arXiv:2605.27589.

[3] VideoCoCo: Code-as-CoT for Physically-Consistent Video Generation via an Agentic Dual-Engine System. arXiv:2607.27380.